Handbook of Semiconductor Manufacturing Technology

Second Edition

Edited by
Robert Doering
Yoshio Nishi

CRC Press
Taylor & Francis Group

Handbook of
Semiconductor Manufacturing Technology

Second Edition

Edited by
Robert Doering
Yoshio Nishi

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2008 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-57444-675-3 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Handbook of semiconductor manufacturing technology / Second Edition edited by Robert Doering and Yoshio Nishi.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-1-57444-675-3
ISBN-10: 1-57444-675-4
1. Semiconductors--Design and construction--Handbooks, manuals, etc. 2. Semiconductor industry--Handbooks, manuals, etc I. Doering, Robert, 1946- II. Nishi, Yoshio, 1940- III. Title.

TK7871.85.H3335 2007
621.3815'2--dc22 2006102599

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Foreword to the Second Edition

In 1958, the year of the invention of the integrated circuit, the price of a single silicon transistor was about $10. Today, it is possible to buy more than 20 million transistors at that price. Not only the transistors but an equal number of passive components and a set of interconnections that permit the devices to function as a dynamic random access memory also can be bought for that price. This cost reduction is unprecedented.

This progress was due to the work of tens of thousands of very capable engineers throughout the world. Every step of the fabrication process, from the preparation of pure silicon materials to the final packaging operations, has been carefully examined, reinvented, and improved. Highly automated equipment have been developed for the manufacturing processes. This book provides an overview of the current status of this work, and takes a look at the expected future growth of the semiconductor industry.

These cost reductions have tremendously expanded the field of electronics. In 1958, the most common electronic products were radio and television sets, and the semiconductor market represented about $218 million. But now semiconductors are used in everything from automobiles to x-ray machines, and its worldwide market exceeds $200 billion. The success story is not yet over as the industry is expected to achieve remarkable growth in the future.

Jack S. Kilby

Preface to the Second Edition

The primary purpose of the second edition of this handbook is to serve as a reference to the practitioners and developers of semiconductor manufacturing technology. Most of the chapters deal with individual process, equipment, material, or manufacturing "control/support/infrastructure" technologies. However, we have supplemented these with a few overview chapters that provide additional background. There is a significant content of graphs, tables, and formulas, which experts in each subfield might find useful. However, unlike a "mostly numbers handbook," we have embedded such compact reference information into a basic description of current and anticipated practice in integrated circuit manufacturing. Although this handbook is not as tutorial as a typical textbook, we have attempted to make each chapter highly readable to nonspecialists in this field. We hope that the book is useful to the scientific community involved in the development and manufacturing of semiconductor products as well as to the students in this field. Note that the book mainly addresses silicon-based manufacturing, although many of the topics are applicable to building devices on other semiconductor substrates. Generous reference lists/bibliographies are also included for the benefit of readers.

Making the best use of the freedom given by editors, the authors have used a breadth of approaches, as well as styles, across the chapters. Some have provided more "historical information," while the others have insisted more on "prognostication." There are chapters that stick pretty closely to the "traditional basics" and others that emphasize the current/future R&D challenges in the area. This flexibility has provided the authors an opportunity to express their personal perspectives and sense of excitement about what is important today and in the near future in this field. Of course, semiconductor manufacturing technology has become so specialized that many of the chapters required multidisciplinary expertise, which in turn has widened perspectives used in the book.

It was definitely challenging to provide up-to-date reference data in a field as broad and rapidly changing as semiconductor manufacturing technology. For example, note the increase in page count found in the National/International Technology Roadmap for Semiconductors editions published from 1992 to 2005. We are glad to publish the second edition of this book, in which we have tried to anticipate what will be most relevant to readers in 2006 and beyond. While most of the authors, who are distinguished technologists in industry and academia, have updated their chapters from the first edition, some others have provided entirely new chapters on topics of current interest in semiconductor manufacturing or its R&D horizon. This continued scaling and evolution of device construction have also resulted in new issues and perspectives, which have influenced the topics as well as their treatments and relationships within this book. Our authors have responded to this challenge with an excellent coordination between and within chapters.

The scenario in semiconductor industry, like when the first edition was published, still remains exciting and we again appreciate this opportunity to share the technical knowledge and experience of the authors/colleagues with the readers.

In addition to thanking the authors, we also like to acknowledge the contributions of Tim Wooldridge, our assistant editor. We have all enjoyed working on this project and hope that the readers too find the book interesting and useful.

Robert Doering and Yoshio Nishi

Editors

Robert Doering is a senior fellow and a technology strategy manager at the Texas Instruments (TI). His previous positions at TI include manager of Future-Factory Strategy, director of Scaled-Technology Integration, and director of the Microelectronics Manufacturing Science and Technology (MMST) Program. The MMST Program was a 5-year R&D effort, funded by the DARPA, the U.S. Air Force, and the TI, which developed a wide range of new technologies for advanced semiconductor manufacturing. The major highlight of the program was the demonstration, in 1993, of sub-3-day cycle time for manufacturing 350-nm CMOS integrated circuits. This was principally enabled by the development of 100% single-wafer processing.

He received a BS degree in physics from the Massachusetts Institute of Technology in 1968 and a PhD in physics from the Michigan State University in 1974. He joined the TI in 1980, after serving several years as the faculty of the physics department at the University of Virginia. His research was based on the nuclear reactions and was highlighted by the discovery of the giant spin–isospin resonance in heavy nuclei in 1973 and by the pioneering experiments in medium-energy heavy-ion reactions in the late 1970s. His early work at the TI was on SRAM, DRAM, and NMOS/CMOS device physics and process flow design. His management responsibilities during the first 10 years at the TI included advanced lithography and plasma etching as well as CMOS and DRAM technology development.

Dr Doering is an IEEE fellow and chair of the Semiconductor Manufacturing Technical Committee of the IEEE Electron Devices Society. He also chairs the National Research Council Board of Assessment for the NIST Electronics and Electrical Engineering Laboratory. He represents the TI on many industry committees, including the Technology Strategy Committee of the Semiconductor Industry Association, the board of directors of the Semiconductor Research Corporation (SRC), the Technical Program Group of the Nanoelectronics Research Corporation, and the Corporate Associates Advisory Committee of the American Institute of Physics. Dr Doering is also one of the two U.S. representatives to the International Roadmap Committee, which governs the International Technology Roadmap for Semiconductors. He has authored/presented over 150 publications and invited papers/talks and has 20 U.S. patents.

Yoshio Nishi is a professor in the Department of Electrical Engineering (research) and also in the Department of Material Science and Engineering at the Stanford University since May 2002. He also serves as the director of the Stanford Nanofabrication Facility of National Nanotechnology Infrastructure Network of the United States and director of research of the Center for Integrated Systems.

He received a BS degree in materials science and a PhD in electronics engineering from Waseda University and the University of Tokyo, respectively.

He researched on semiconductor device physics and silicon interfaces in the Toshiba R&D, which resulted in the discovery of ESR PB Center at SiO2–Si interface, the first 256-bit MNOS nonvolatile RAM, SOS 16-bit microprocessor, and the world's first 1-MB CMOS DRAM.

In 1986 he joined Hewlett-Packard as the director of the Silicon Process Lab, and then established the ULSI Research Lab.

Dr Nishi joined the TI, Inc. in 1995 as the senior vice president and the director of R&D for the semiconductor group, implemented a new R&D model for silicon technology development, and established the Kilby Center.

Since May 2002, he became a faculty member at the Stanford University, and his research interest covers nanoelectronic devices and materials including a metal gate/high-k MOS, a device layer transfer for 3D integration, nanowire devices, and resistance change nonvolatile memory materials and devices. He has published more than 200 papers including conference proceedings, and he coauthored/edited nine books. He holds more than 70 patents in the United States and Japan.

During the period of 1995–2002, he served as a board member of the SRC and the International Sematech, the NNI panel, the MARCO governing council, and other boards. Currently, he serves as an affiliated member of the Science Council of Japan.

Dr Nishi is a fellow of the IEEE, a member of the Japan Society of Applied Physics and the Electrochemical Society. His recent awards include the 1995 IEEE Jack Morton Award and the 2002 IEEE Robert Noyce Medal.

Contributors

Michael Ameen
Axcelis Technologies, Inc.
Beverly, Massachusetts

Nick Atchison
Multigig, Inc.
Scotts Valley, California

Sanjay Banerjee
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, Texas

Gabriel G. Barna
Process Development and Control
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Robert Baumann
Component Reliability Group
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Ivan Berry
Axcelis Technologies, Inc.
Beverly, Massachusetts

Duane S. Boning
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts

Louis Breaux
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Stephanie Watts Butler
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Jeff Byers
KLA-Tencor
Austin, Texas

Andreas Cangellaris
Department of Electrical and Computer Engineering
The University of Illinois
Urbana, Illinois

G. K. Celler
Soitec USA
Peabody, Massachusetts

Mei Chang
Applied Materials, Inc.
Santa Clara, California

Walter Class
Axcelis Technologies, Inc.
Beverly, Massachusetts

C. Rinn Cleavelin
Silicon Technology Development
Texas Instruments, Inc.
Austin, Texas

Sean Collins
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Luigi Colombo
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Will Conley
Freescale Semiconductor, Inc.
Austin, Texas

Sorin Cristoloveanu
Institute of Microelectronics, Electromagnetism and Photonics
Grenoble, France

Francois M. d'Heurle
IBM Thomas J. Watson Research Center
Yorktown Heights, New York

Vallabh H. Dhudshia
SafeFab Solutions
Plano, Texas

Alain C. Diebold
SEMATECH
Austin, Texas

Girish A. Dixit
Novellus Systems, Inc.
San Jose, Califonia

Simon Fang
United Microelectronics Corporation
Hsin-Chu City, Taiwan

Leonard Foster
Facilities Department
Texas Instruments, Inc.
Dallas, Texas

Gene E. Fuller
Strategic Lithography Services
Punta Gorda, Florida

Glenn W. Gale
SEZ AG
Villach, Austria

César M. Garza
Freescale Semiconductor, Inc.
Austin, Texas

Hans-Joachim Gossmann
Axcelis Technologies, Inc.
Beverly, Massachusetts
and
Advanced Micro Devices
Hopewell Junction, New York

Gautum Grover
Cabot Corporation
Aurora, Illinois

John R. Hauser
Electrical and Computer Engineering Department
North Carolina State University
Raleigh, North Carolina

Robert H. Havemann
Novellus Systems, Inc.
San Jose, Califonia

Howard Huff
SEMATECH
Austin, Texas

G. Dan Hutcheson
VLSI Research, Inc.
Santa Clara, California

Frederick W. Kern, Jr.
Hitachi Global Storage Technologies
San Jose, California

Brian K. Kirkpatrick
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Vincent Korthuis
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Michael Lamson
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Christian Lavoie
IBM Thomas J. Watson Research Center
Yorktown Heights, New York
and
Département de Génie Physique
École Polytechnique de Montréal
Montréal, Canada

Wen Lin
Consultant
Allentown, Pennsylvania

Erdogan Madenci
Department of Aerospace and Mechanical Engineering
University of Arizona
Tucson, Arizona

Andrew J. McKerrow
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

J. W. McPherson
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Mohammed J. Meziani
Department of Chemistry and
Laboratory for Emerging Materials and Technology
Clemson University
Clemson, South Carolina

Hiro Niimi
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

E. T. Ogawa
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Sylvia Pas
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Pankaj Pathak
Department of Chemistry and Laboratory for Emerging Materials and Technology
Clemson University
Clemson, South Carolina

Devadas Pillai
Intel Corporation
Chandler, Arizona

Shahid Rauf
Freescale Semiconductor, Inc.
Austin, Texas

Jonathon Reid
Novellus, Inc.
San Jose, California

Syed A. Rizvi
Nanotechnology Education and Consulting Services
San Jose, California

Ron Ross
Texas Instruments, Inc.
Santa Cruz, California

Stephen M. Rossnagel
IBM Thomas J. Watson Research Center
Yorktown Heights, New York

Leonard Rubin
Axcelis Technologies, Inc.
Beverly, Massachusetts

Dieter K. Schroder
Electrical Engineering Department
Arizona State University
Tempe, Arizona

Bruno W. Schueler
Revera Inc.
Sunnyvale, California

Thomas E. Seidel
AIXTRON, Inc.
Sunnyvale, California

Thomas Shaffner
National Institute of Standards and Technology
Gaithersburg, Maryland

Gregory B. Shinn
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Terry Sparks
Freescale Semiconductor, Inc.
Austin, Texas

Greg S. Strossman
XPS/ESCA and TOF-SIMS Services
Evans Analytical Group
Sunnyvale, California

Ya-Ping Sun
Department of Chemistry and Laboratory for Emerging Materials and Technology
Clemson University
Clemson, South Carolina

P. J. Timans
Mattson Technology
Fremont, California

Ting Y. Tsui
Silicon Technology Development
Texas Instruments, Inc.
Dallas, Texas

Peter L. G. Ventzek
Freescale Semiconductor, Inc.
Austin, Texas

Brad VanEck
SEMATECH
Austin, Texas

Eric M. Vogel
Department of Electrical Engineering
University of Texas at Dallas
Dallas, Texas

Lawrence C. Wagner
Semiconductor Quality Department
Texas Instruments, Inc.
Dallas, Texas

Samuel C. Wood
Responsive Learning Technologies
Los Altos, California

Li-Qun Xia
Applied Materials, Inc.
Santa Clara, California

Shi-Li Zhang
School of Information and Communication Technology
Royal Institute of Technology
Stockholm, Sweden
and
School of Microelectronics
Fudan University
Shanghai, China

Contents

1 Introduction to Semiconductor Devices *John R. Hauser* **1**-1

2 Overview of Interconnect—Copper and Low-*K* Integration *Girish A. Dixit and Robert H. Havemann* **2**-1

3 Silicon Materials *Wen Lin and Howard Huff* **3**-1

4 SOI Materials and Devices *Sorin Cristoloveanu and George K. Celler* **4**-1

5 Surface Preparation *Glenn W. Gale, Brian K. Kirkpatrick, and Frederick W. Kern, Jr.* **5**-1

6 Supercritical Carbon Dioxide in Semiconductor Cleaning *Mohammed J. Meziani, Pankaj Pathak, and Ya-Ping Sun* **6**-1

7 Ion Implantation *Michael Ameen, Ivan Berry, Walter Class, Hans-Joachim Gossmann, and Leonard Rubin* **7**-1

8 Dopant Diffusion *Sanjay Banerjee* **8**-1

9 Oxidation and Gate Dielectrics *C. Rinn Cleavelin, Luigi Colombo, Hiro Niimi, Sylvia Pas, and Eric M. Vogel* **9**-1

10 Silicides *Christian Lavoie, Francois M. d'Heurle and Shi-Li Zhang* **10**-1

11 Rapid Thermal Processing *P.J. Timans* **11**-1

12 Low-*K* Dielectrics *Ting Y. Tsui and Andrew J. McKerrow* **12**-1

13 Chemical Vapor Deposition *Li-Qun Xia and Mei Chang* **13**-1

14 Atomic Layer Deposition *Thomas E. Seidel* **14**-1

15 Physical Vapor Deposition *Stephen M. Rossnagel* **15**-1

16 Damascene Copper Electroplating *Jonathan Reid* **16**-1

17 Chemical–Mechanical Polishing *Gregory B. Shinn, Vincent Korthuis, Gautum Grover, Simon Fang, and Duane S. Boning* **17**-1

18 Optical Lithography *Gene E. Fuller* **18**-1

19 Photoresist Materials and Processing *César M. Garza, Will Conley, and Jeff Byers* **19**-1

20 Photomask Fabrication *Syed A. Rizvi and Sylvia Pas* **20**-1

21 Plasma Etch *Peter L.G. Ventzek, Shahid Rauf, and Terry Sparks* **21**-1

22 Equipment Reliability *Vallabh H. Dhudshia* **22**-1

23 Overview of Process Control *Stephanie Watts Butler* **23**-1

24 In-Line Metrology *Alain C. Diebold* **24**-1

25 In-Situ Metrology *Gabriel G. Barna and Brad VanEck* **25**-1

26 Yield Modeling *Ron Ross and Nick Atchison* **26**-1

27 Yield Management *Louis Breaux and Sean Collins* **27**-1

28 Electrical, Physical, and Chemical Characterization *Dieter K. Schroder, Bruno W. Schueler, Thomas Shaffner, and Greg S. Strossman* **28**-1

29 Failure Analysis *Lawrence C. Wagner* **29**-1

30 Reliability Physics and Engineering *J.W. McPherson and E.T. Ogawa* **30**-1

31 Effects of Terrestrial Radiation on Integrated Circuits *Robert Baumann* **31**-1

32 Integrated-Circuit Packaging *Michael Lamson, Andreas Cangellaris, and Erdogan Madenci* **32**-1

33 300 mm Wafer Fab Logistics and Automated Material Handling Systems *Leonard Foster and Devadas Pillai* **33**-1

34 Factory Modeling *Samuel C. Wood* **34**-1

35 Economics of Semiconductor Manufacturing *G. Dan Hutcheson* **35**-1

Appendix A: Physical Constants **A**-1
Appendix B: Units Conversion **B**-1
Appendix C: Standards Commonly Used in Semiconductor Manufacturing **C**-1
Appendix D: Acronyms **D**-1

Index **I**-1

1

Introduction to Semiconductor Devices

1.1 Introduction **1**-1
1.2 Overview of MOS Device Characteristics **1**-3
1.3 MOSFET Device Scaling **1**-8
Scaling Rules • Performance of Scaled Devices
1.4 Manufacturing Issues and Challenges **1**-22
MOSFET Gate Stack Issues • Channel Doping Issues • Source/Drain Contact Issues • Substrate and Isolation Issues • Thermal Budget Issues
1.5 Advanced MOS Device Concepts **1**-44
SOI Substrates and Devices • Multiple Gate MOS Devices • Transport Enhanced MOS Devices • MOSFETS with Other Semiconductors • Advanced Semiconductor Device Concepts
1.6 Conclusions **1**-53
References **1**-53

John R. Hauser
North Carolina State University

1.1 Introduction

The silicon metal oxide semiconductor field effect transistor (MOSFET) has emerged as the ubiquitous active element for silicon very large scale integration (VLSI) integrated circuits. The competitive drive for improved performance and cost reduction has resulted in the scaling of circuit elements to ever-smaller dimensions. Within the last 35 years, MOSFET dimensions have shrunk from a gate length of 5 μm in the early 1970s to 45 nm today, and are forecast to reach less than 10 nm at the end of the projected shrink path in about 2020. While this process has been driven by market place competition with operating parameters determined by products, manufacturing technology innovations that have not necessarily followed such a consistent path have enabled it. This treatise briefly examines metal oxide semiconductor (MOS) device characteristics and elucidates important future issues which semiconductor technologists face as they attempt to continue the rate of progress to the identified terminus of the technology shrink path in about 2020.

In the early days of semiconductor device development (the 1950s), the bipolar junction transistor was the dominant semiconductor device. As large-scale integration of devices developed in the 1960s, the MOSFET became the preferred device type and has eventually grown to dominate the use of semiconductor devices in integrated circuits (ICs). This has been predominantly due to the development of complementary MOS devices (CMOS), where digital logic circuits can be formed that exhibit extremely low power dissipation in either of the two logic states. Complementary MOS is not only a device technology, but also a logic circuit technology that dominates the IC world because of the advantages of very low power dissipation over other forms of semiconductor circuits. Thus, over time in

the general scheme of ICs, bipolar semiconductor devices have come to be used in only special applications. Because of this dominance of MOS devices in large-scale ICs, only the MOSFET will be reviewed in this discussion.

In the late 1960s, the semiconductor industry emerged from the era of wet processing, contact printing, and negative resist at approximately 10 μm minimum gate lengths to face considerable problems related to particulate reduction, dry processing, and projection printing. Positive resist overcame the resist-swelling problem and provided improved resolution, but at the cost of particle generation caused by brittle resist and railed wafer handling equipment in use at that time. Whole wafer projection printing improved the resolution and yield, but required the use of larger wafers to offset the additional capital cost. Dry processing was the banner developmental thrust of the period. Plasma processing, initially conducted in a pancake reactor between opposing electrodes dramatically increased yield, but required a trained artisan to achieve uniformity and throughput. Sputter metal deposition replaced evaporation that had earlier produced substrate stress voiding problems. Wafer size was increased to offset the cost of the more sophisticated and expensive process equipment. Dynamic random access memory factories were the workhorses to develop and prove out the next technology generation. The MOSFET began to emerge in the 1970s as the device technology for VLSI, although large-scale integration (LSI) bipolar technology persisted longer than many forecast.

In the early 1980s, Japanese semiconductor manufacturers seized manufacturing technology leadership with major capital commitments, dramatically increasing manufacturing yield and factory efficiency, to capture a major share of the Dynamic random access memories (DRAM) market. Quality became a major issue when it was reported that quality levels of Japanese memories were consistently and substantially better than American manufacturers. As the cost of manufacturing equipment development escalated, a transition began with the emergence of dedicated manufacturing equipment companies, reducing the value of proprietary process development and enhancing the value of product definition and circuit design. Major semiconductor companies initially pressured these equipment vendors to "customize" each production tool to their proprietary specifications, inhibiting reduced costs of capital equipment. Japan became a major supplier of semiconductor manufacturing equipment, further exacerbating problems for the U.S. semiconductor industry. This scenario produced a strategic inflection point in the IC production: U.S. vendors of DRAM suffered financial losses and many ultimately exited the mass memory market.

This situation spawned cooperation among the major U.S. semiconductor manufacturers, resulting in the establishment of the Semiconductor Research Corp. (SRC) in 1981 and of SEMATECH in 1988. Both of these pursued the concept of industrial collaboration in semiconductor research and the concept of equipment "cost of ownership", leading to significant collaboration with IC manufacturers and equipment vendors. Wafer size was again increased to reduce the cost of IC manufacturing. Meanwhile, the relentless march of the technology to smaller feature size continued. This was made possible by step-and-repeat projection printing, reduced generation of particulates and single wafer processing replacing many batch processing steps. Microprocessor technology emerged with dynamic memory as a separately addressed technology requirement. Subsequently, the microprocessor manufacturing technology led the race toward smaller device structures and higher complexity on the chip. Dynamic memory continues to advance with smaller memory cell size and ever increasing memory size.

Manufacturing equipment automation, cleanliness (for particle reduction) and efficient factory management through cost of ownership reductions have become the overriding issues to be addressed. Yield was increased in the 1980s and 1990s three- and fourfold to the levels unheard of in the early years of IC manufacturing. The collaborative work of SEMATECH and the U.S. semiconductor industry led to an initial "roadmap" of semiconductor technology in 1992 and to the subsequent revisions approximately every 2 years. Significantly aided by SRC and SEMATECH, the U.S. regained the technology and market leadership by the early 2000s in the world semiconductor markets.

Since the 1990s, enhanced collaboration among semiconductor companies has occurred to continually recognize and address technology problems associated with the continuing decrease in device feature size. Old issues have persisted and many new ones have arisen. Among the most important issues are: (1)

the increasing capital cost of manufacturing plants, (2) the increased difficulty of lithographic processes as feature sizes decrease, (3) the increasing cross-talk and capacitive loading on-chip as frequency of operation is increased, (4) power dissipation problems, (5) fundamental device limits as feature sizes approach the nanometer range, and (6) the need for enhanced metrology and test equipment. In addition, system and circuit design have presented additional sets of problems that have been separately addressed by a computer-aided design (CAD) industry that arose in a similar fashion to the manufacturing equipment industry.

Today at device feature sizes below 50 nm, the semiconductor industry faces unprecedented problems with the task of continuing affordable feature size scaling with the continued improvement of the manufacturing processes. As will be subsequently discussed, fundamental device scaling issues have changed at around the 100–50 nm feature size and many new materials and processes are being required to continue scaling to the ultimate limits imposed by semiconductor device physics. Key areas for advancement of the MOSFET include advanced gate dielectrics, gate contacts, source/drain structures and contacts to the source/drain, and the possible replacement of bulk CMOS with silicon on insulator (SOI) CMOS or eventually with more complex three-dimensional device structures that might perform improved circuit functions. To implement these new materials and smaller device dimensions, new manufacturing methods and equipment will be required with tighter tolerances, lower manufacturing cost of ownership, and increased process control. Improved predictability from improved technology computer-aided design (TCAD) is necessary to avoid the expense of process development on the manufacturing line. Thus, the battle continues to ever increase device density and lower costs through smaller device dimensions as the infusion of semiconductor devices continues into a still rapidly growing electronics market.

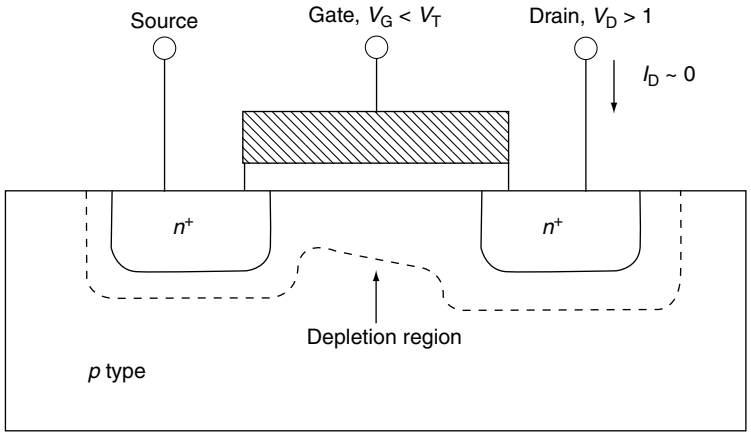
1.2 Overview of MOS Device Characteristics

The MOSFET, is the predominant semiconductor device, in large-scale ICs. The basic physics behind the device can be understood with reference to Figure 1.1, which shows a cross-section of the basic components of an *n*-channel device. In the MOS device, a voltage applied to the gate (V_G in Figure 1.1) controls the current flow between the drain and the source. The physical mechanism of control is through controlling the mobile charge density of a channel of electrons near the interface between the silicon and an insulator. The gate-to-semiconductor structure acts in many ways like a capacitor with the channel charge controlled by the gate-to-source voltage. One difference with a simple capacitor is the existence of a threshold voltage V_T, which is required at the gate to establish the onset of a conductive channel. This is due to the background doping density of the substrate, which for an *n*-channel device as shown is of opposite conductivity type or *p*-type. The applied gate voltage must establish a depletion layer of some width with an accompanying voltage before the conductive channel of electrons, begins to dominate the capacitor charge.

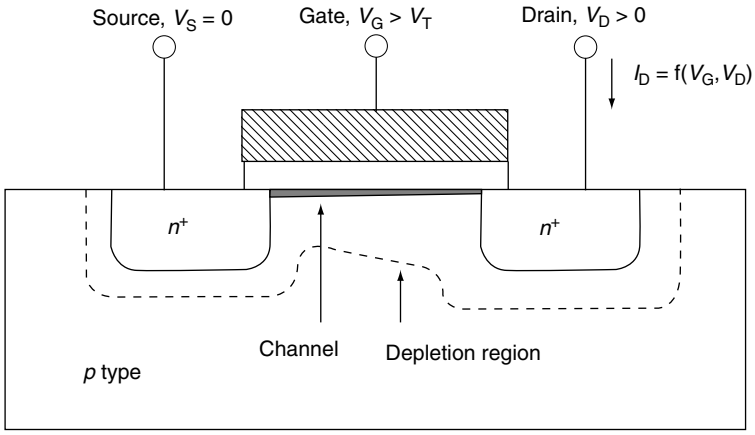
A typical capacitance–voltage characteristic for an *n*-MOS transistor is shown in Figure 1.2 for the case of source and drain tied together. For voltages less than V_T in the figure, only a depletion region (with positive charge density) exists or an accumulation layer of holes (positive charge) exists for a large negative gate voltage (with negative gate charge). The minimum in the capacitance corresponds to a maximum value of the depletion layer established in the semiconductor substrate. The width of this maximum depletion layer can be expressed as:

$$W_{dm} = \sqrt{\frac{4\varepsilon_{si} kT \ln(N_a/n_i)}{q^2 N_a}} \tag{1.1}$$

where N_a is the acceptor doping density in the semiconductor and n_i is the intrinsic carrier density. This also occurs theoretically when the surface potential has achieved the value:



(a) "off" state, $V_G < V_T$, no channel exists



(b) "on" state, $V_G > V_T$, conductive channel exists

FIGURE 1.1 Schematic of basic *n*-channel metal oxide semiconductor field effect transistor (MOSFET). (a) In "off" state, (b) In "on" state.

$$\psi_s = 2(kT/q)\ln(N_a/n_i) = 2\phi_B \tag{1.2}$$

The rapidly increasing capacitance curve for voltages above V_T is indicative of the rapid establishment of an inversion layer above the depletion layer, which in this case, is a conductive channel of electrons. The heavily doped n^+ source/drain regions shown in Figure 1.1 are used to make "ohmic" contact to the conductive channel so that a voltage difference between the source and the drain will result in current flow from the positive voltage at the drain terminal to the negative voltage at the source. In the "off" stage as illustrated in Figure 1.1A, the drain current is very small (ideally zero) and in the "on" state as illustrated in Figure 1.1B, the drain current is a function of both V_G and V_D. The larger the gate voltage, the larger will be the density of conduction channel electrons and the larger will be the device drain current. Ideally, for gate voltages significantly above the threshold voltage, the gate-to-channel looks

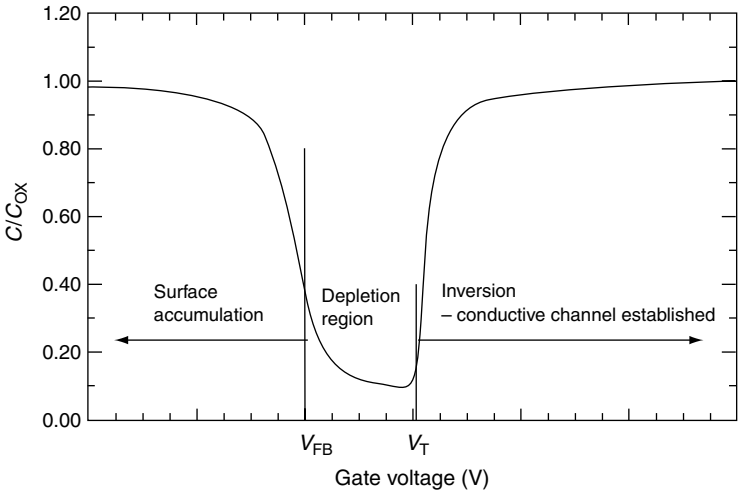


FIGURE 1.2 Gate *C–V* curve for MOSFET with drain and source at same voltage.

like a capacitor of dielectric thickness equal to the gate oxide thickness and the capacitance approaches a constant as seen by the curve in Figure 1.2 for large positive gate voltages.

A typical graph of I_D vs. V_D for several gate voltages is shown in Figure 1.3. This is for a long-channel device that shows little of the "short-channel" effects to be subsequently discussed. The I–V characteristic exhibits three distinct regions of operation: (1) cutoff or subthreshold, where $V_G < V_T$, (2) the triode region, where $V_D < V_G - V_T$, and (3) the saturation region, where $V_D > V_G - V_T$ and the current is approximately independent of drain voltage as shown in Figure 1.3 for large drain voltages. Ideally in the subthreshold region the current would be zero, but the conductive channel does not go abruptly to zero at the threshold voltage, so an exponentially decreasing current exists in the subthreshold region due to an exponentially decreasing inversion charge density. This is best illustrated by the typical log plot of

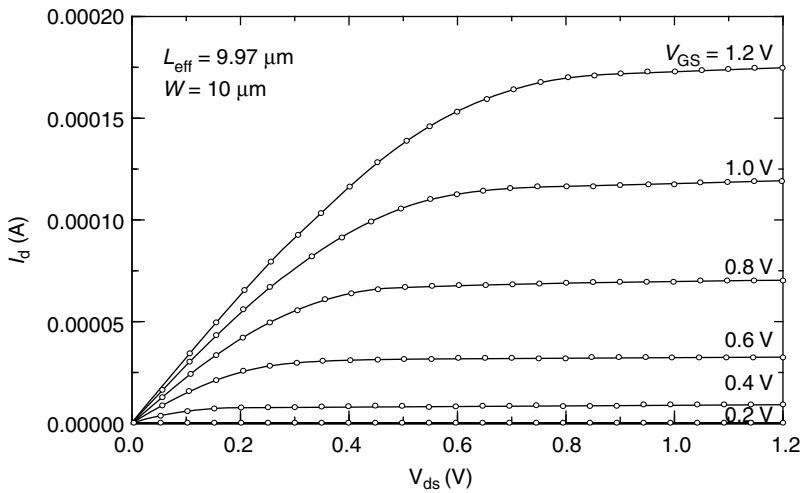


FIGURE 1.3 Typical I_D–V_D characteristic for long *n*-channel metal oxide semiconductor field effect transistor at constant gate voltages.

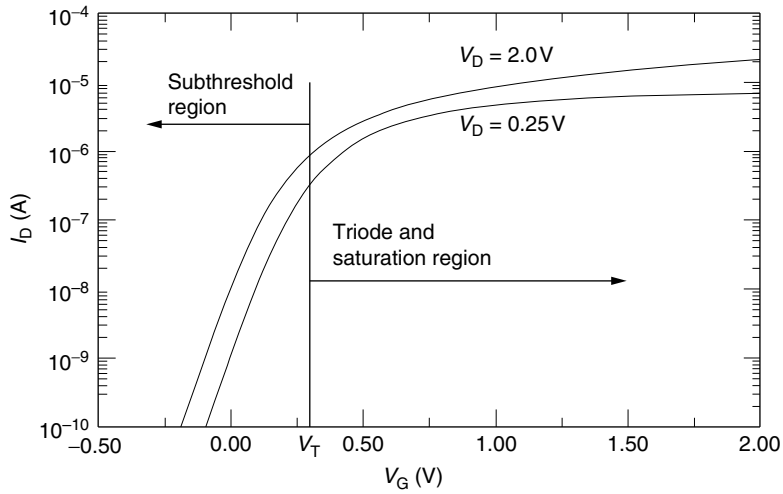


FIGURE 1.4 Typical I_D–V_G MOSFET characteristic in the subthreshold region.

I_D vs. V_G at two drain voltages as shown in Figure 1.4. As can be seen in the figure, the current is approximately linear on the log plot over the voltage region below the threshold voltage.

The standard first-order model for MOSFET terminal current is the set of equations [1]:

$$I_D = \begin{cases} \text{(a) Subthreshold region} & I_o \exp(-q(V_T - V_G)/mkT) \quad \text{for } V_G < V_T \\ \text{(b) Triode region} & \mu_n C_{ox}(W/L)(V_G - V_T - V_D/2)V_D \quad \text{for } V_G > V_T \text{ and } V_D < V_G - V_T \\ \text{(c) Saturation region} & \mu_n C_{ox}(W/2L)(V_G - V_T)^2 \quad \text{for } V_G > V_T \text{ and } V_D > V_G - V_T \end{cases} \tag{1.3}$$

These are approximate equations that help to define the three major regions of operation as (a) subthreshold region, (b) triode region (or linear region for small values of V_D), and (c) saturation region. The equation for the subthreshold region must be matched to the conduction region equation in a manner such that the value and first derivative of the current are continuous. The equations are used here to illustrate the major dependences on the device parameters. The device and structural dependences occur through the effective channel length L, the channel width W, the channel mobility μ_n, and the inversion layer capacitance C_{ox}. An additional parameter frequently used to characterize the subthreshold region is the current slope factor:

$$S = m(kT/q)\ln(10) \simeq (kT/q)\ln(10)(1 + \varepsilon_{si} t_{ox}/\varepsilon_{ox} W_{dm}) \tag{1.4}$$

In this equation, a model for the ideality factor m is also given in terms of the oxide thickness and the maximum depletion layer width. This value is frequently compared to an ideal subthreshold slope factor of $(kT/q)\ \ln(10) = 60$ mV/decade at room temperature.

For short-channel devices, additional parameters are needed to provide even a first-order description of device current. A typical set of current characteristics for a short-channel MOSFET are shown in Figure 1.5. In this particular case, it is the characteristic of an n-channel device with an oxide thickness of approximately 1.7 nm, a channel length of approximately 97 nm, and a channel width of 10 μm. What constitutes a short-channel device is not readily defined simply in terms of the effective channel length. As devices have been scaled to ever smaller dimensions over the years, the concept of a short-channel

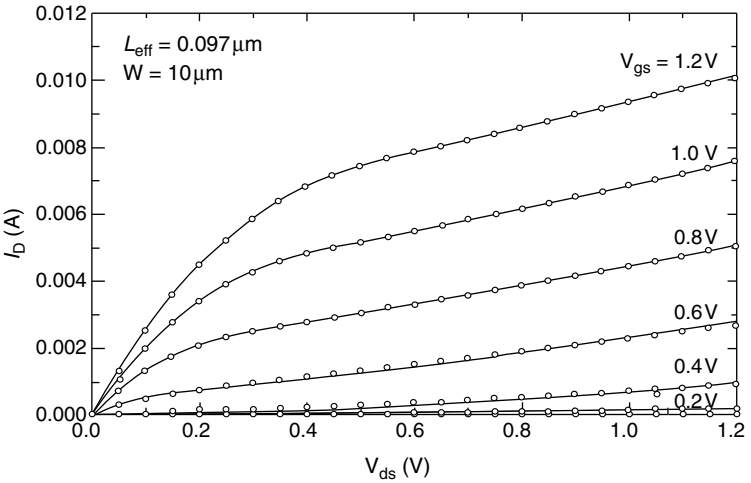


FIGURE 1.5 Typical I_D–V_D characteristic for short *n*-channel MOSFET at constant gate voltages.

device has been pushed to ever smaller dimensions. This scaling is discussed in a subsequent section. However, devices that can be characterized as short-channel devices exhibit certain important deviations from the ideal device equations as described by the set of Equation 1.3. One important feature is the lack of a clear current saturation region for large drain voltages as can be seen in comparing Figure 1.3 and Figure 1.5. This lack of ideal current saturation is known to be due primarily to two physical effects. One is the channel-length modulation, whereby an increase in drain voltage reduces the effective channel length. As can be seen from Equation 1.3, a decrease in L will cause a resulting increase in the device current even for the equation describing the current saturation region. The second effect is a dependence of device threshold voltage on drain voltage with the threshold voltage decreasing with increasing drain voltage. Again in Equation 1.3, we can see that this will result in an increasing device current. This decrease in threshold voltage is due to an incomplete shielding of the channel from the drain voltage and depends strongly on the three parameters: oxide thickness, maximum depletion layer thickness, and effective channel length. This is a two-dimensional potential feedback effect from the drain to the channel. This is discussed in more detail in a subsequent section, but for a typical drain current characteristic, is typically described by a so called drain-induced barrier lowering (DIBL) factor.

The channel-length modulation effect is frequently included in the basic device model by the use of a factor $(1+\lambda_n V_D)$, which is used to multiply Equation 1.3 in the triode and saturation regions. A first order model for DIBL can be taken as $\Delta V_T = \sigma V_D$ in all regions of operation. These are the two most important modifications of the basic device equations needed for short-channel effects. Both of these modifications will result in a finite slope on the *I–V* characteristic at the large drain voltages in the classical current saturation region. Most of the finite slope seen in Figure 1.5 can be accounted for by a DIBL effect (with $\sigma \simeq 0.22$).

A key factor in the dominance of MOSFETs is the ability to produce complementary *n*- and *p*-channel transistors with similar device characteristics. The figures given so far have been for *n*-channel devices in which a positive gate voltage is required to establish a conductive channel and current flows from the most positive of the source/drain contacts to the most negative. A discussion of the *p*-channel devices would be essentially the same, except for the reversal of all operating voltages and current directions. A negative voltage at the gate first depletes an *n*-type substrate and then establishes a conductive channel of holes (positive charge) after the threshold voltage is exceeded. Positive current can then flows from the most positive of the source/drain contacts to the most negative with the most negative contact identified as the drain contact. Figures showing the *I–V* characteristics of the *p*-channel devices would be very similar to Figure 1.3 through Figure 1.5 with the current and voltage directions reversed. One major

difference should, however, be noted and this relates to the mobility parameter (which would be μ_p in Equation 1.3), which is about a factor of 2.0–2.5 times smaller for holes than for electrons. However, the current supplied by a *p*-MOS can be equal to that of an *n*-MOS by increasing the (W/L) ratio for the *p*-MOS to compensate for the lower hole mobility. It is frequently assumed that (W/L) for the *p*-MOS devices is approximately 2.5 times that of the *n*-MOS devices, although logic gates may require other ratios of the device dimensions in order to achieve desired switching speeds.

1.3 MOSFET Device Scaling

The primary driving factor behind the exponential increase over the past 40 years of IC functionality, commonly referred to as Moore's law, has been the ability to continually scale MOS devices to ever smaller dimensions. This is illustrated in Figure 1.6, which shows a series of experimental devices (production and experimental devices) spanning the major technology nodes from 130 to 32 nm, and the expected production times from 2001 to 2009 [2]. Figure 1.7 shows a cross-sectional drawing of a typical MOS device drawn approximately to scale in both the vertical and horizontal dimensions for devices in the range of the 130- to 32-nm nodes. Critical dimensions on the figure are the physical gate length L_P, the gate oxide thickness t_{ox}, and the effective channel length L_{eff} (same as L in Equation 1.3). Of particular note is the very thin nature of the gate oxide when compared with other device dimensions such as the effective channel length. This structure shows a typical channel contacting structure consisting of the source/drain extension junction (depth X_J) and a deeper drain contact junction (depth X_{JC}). Devices typically have a silicide layer extending nearly to the edge of the drain extension layer and a metal source/drain and gate contact layer, which are not shown in this simplified drawing. These layers can be seen in the actual cross-sectional views of Figure 1.6. Also not shown in Figure 1.7 are sub-channel doping layers under the gate in the region of L_{eff} for controlling threshold voltage, punch-through and DIBL. These are discussed in more detail later.

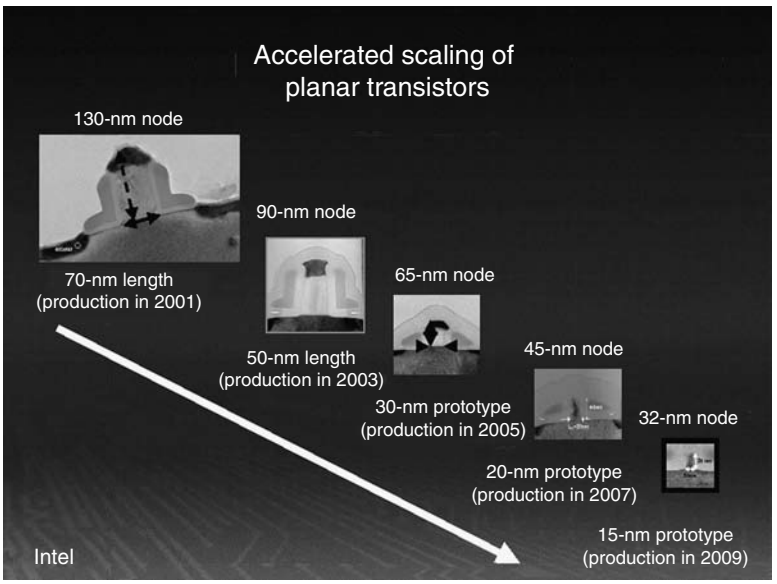


FIGURE 1.6 Illustration of device scaling from the 130-nm node to the 32-nm node. (From Marcyk, G., INTEL Corp., ftp://download.intel.com/technology/silicon/Marcyk_tri_gate_0902.pdf)

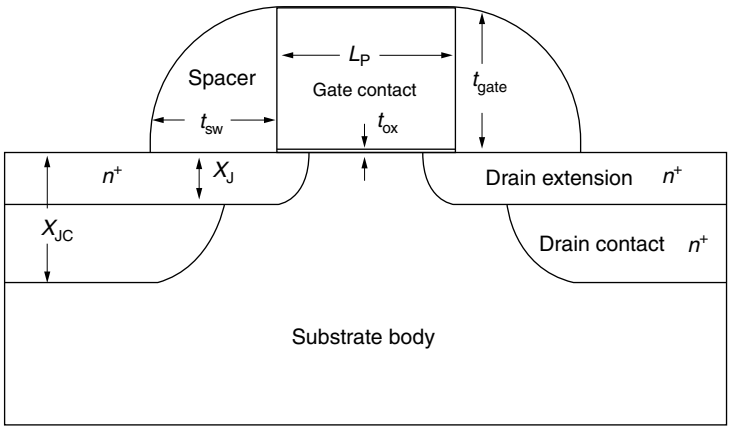


FIGURE 1.7 Metal oxide semiconductor device cross-section drawn approximately to scale in vertical and horizontal directions.

1.3.1 Scaling Rules

Since first published in 1992, the International Technology Roadmap for Semiconductors (ITRS) [3] has been the benchmark document for guiding the scaling of MOS devices. Each edition of this document has projected future scaling trends and future device and IC system performance for the next 15 years. The 2004 update [4] projects development to the year 2019 when there is a considerable concern that the end of any practical CMOS scaling will have been reached due to fundamental material and device limits. Several of these limits and problems will be subsequently discussed. Figure 1.8 shows some important projected scaling results from both the 1997 and the 2004 ITRS documents. The figure shows several device parameters in terms of major technology generations and the year (or expected year) of its

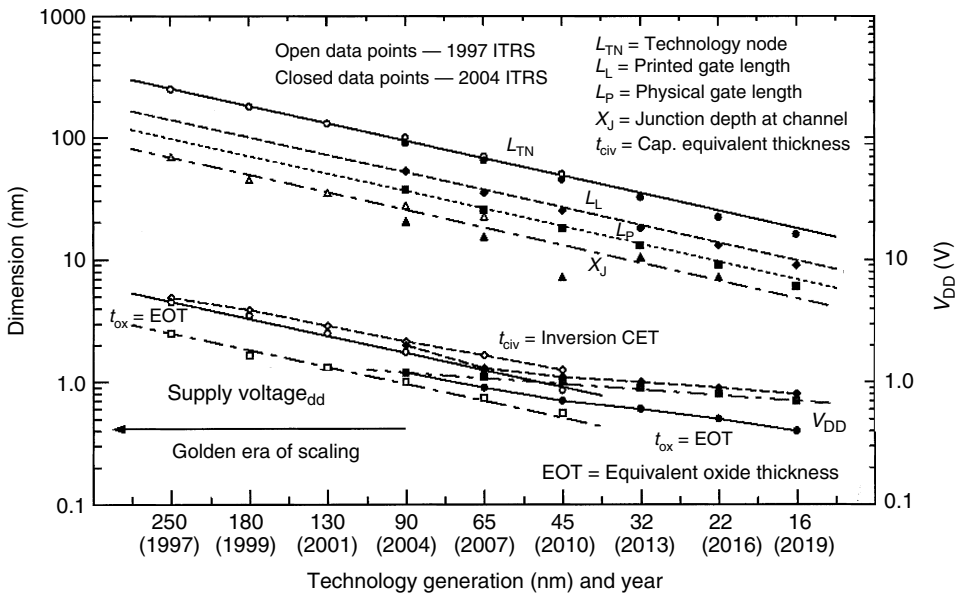


FIGURE 1.8 Dimensions of several important metal oxide semiconductor device parameters as projected for scaled devices.

introduction into production. The technology generation nodes need some discussion. In the 1997 ITRS, the nodes were listed as 250, 180, 130, 100, 70, and 50 nm. In the 2004 ITRS, the last three of these nodes have been changed to 90, 65, and 45 and the additional nodes of 32, 22, and 16 are included. One of the intents of the major technology nodes is to identify device scaling that will result in an increase in device packing density by a factor of 2 (a doubling of device density each generation along Moore's curve). Ideally, then each technology generation represents a shrink in linear device dimension by a factor of $1/\sqrt{2} = 0.707$. With this in mind the 90-nm dimension is considerably closer to an ideal shrink than was the original 100 nm specification. Thus, the 1997 data points are plotted using the 2004 node identifications to be consistent with the 2004 document. Also to be noted from the expected dates of introduction, the time between the major nodes is not a constant time interval. For the 250, 180, and 130 nodes, the time interval is only 2 years, while for the remaining nodes the projected time interval is 3 years. This indicates some of the expected difficulty in achieving the expected shrinks beyond the 130 nm node.

Several lengths are shown in the graph from the 2004 ITRS. First, the upper curve labeled L_{TN} corresponds to the major technology node, which is identified with the expected half-pitch of densely packed DRAMS circuits. For the critical MOS device dimensions, especially of "high performance" (HP) logic devices, other critical dimensions are defined as L_L, the lithographically printed gate length; and L_P, the physical gate length after etching of the gate contact material. Also shown in the figure are projections for the depth of the extension junction used to contact the conductive channel. As can be seen by the approximate straight lines on the log scale of Figure 1.8, all of these dimensions are projected to scale as some constant fraction of the technology node dimension. It should be noted that, the 2004 ITRS did not distinguish between the technology node and the physical device dimension, so there was some confusion about these dimensions in the earlier ITRS documents. It should be noted that, the 2004 ITRS develops different scenario for HP logic circuits and for low power circuits. While the line width parameters are common for different applications, other parameters such as dielectric thickness and supply voltage, which are included in Figure 1.8 are application dependent. In general, parameters discussed in this chapter are for the HP logic circuits, since these applications place the most severe constraints on the device performance and will probably be the most difficult one to achieve with scaled devices.

Other very key device parameters are the gate dielectric thickness and the logic level power supply voltage. First, considering the supply voltage V_{DD}, in the 2004 ITRS, it was projected to decrease to about 0.5 V by the 45-nm node or by the end of the 2004 projected roadmap. The 2004 ITRS provides a very different scenario with the 45-nm node having a supply voltage of 1.0 V and the value decreasing to 0.7 V at the 16-nm node or the end of the roadmap. This is a major rethinking of the voltage limits for scaled CMOS devices. Along with this has been a major change in the projected decrease in dielectric thickness. In the 2004 ITRS, the oxide thickness was projected to reach about 0.9 nm (actually listed as <1 nm) at the 45-nm node. Actually the 2004 ITRS projects even thinner t_{ox} or equivalent oxide thickness (EOT) values than did the 1997 ITRS. In this case, EOT in recognition that the gate dielectric will most likely not be a pure oxide, but some enhanced oxide for reduced gate leakage. This is explored in more detail in a subsequent section. An additional parameter introduced in the 2004 ITRS is the "equivalent electrical thickness in inversion" (t_{civ}) as also plotted in Figure 1.8. This is also labeled "inversion CET" in Figure 1.8 for inversion capacitance equivalent thickness (CET). It should be noted that, the supply voltage and oxide thickness values used here are for so called HP devices such as would be used in the state-of-the-art microprocessors. Other MOS devices needed for low power applications would not be scaled as aggressively, but would be optimized for other considerations. The HP device scaling is emphasized here as it typically represents the most severe scaling constraint in terms of material properties and device dimensions.

A brief discussion is needed for the inversion CET capacitance as this is a critical parameter for MOS device operation. Referring back to Figure 1.1, the conductive channel induced by the gate voltage into the semiconductor substrate acts much like the plate of a capacitor with the gate electrode as one plate and the conductive channel as the other electrode. Ideally, the capacitance of the gate-to-channel is that of a

capacitor with dielectric constant equal to that of the gate insulator and thickness of the gate insulator t_{ox} or EOT. For such an ideal case, the CET would equal the EOT. However, this is not quite correct, especially when the dimensions get very small. First, if the gate is polysilicon with a finite doping density, there is a finite depletion layer that exists in the polysilicon gate. This essentially adds to the equivalent dielectric thickness of the gate-to-channel capacitor structure. The ITRS projections for the ultimately scaled devices assume that this additional thickness will be eliminated by the use of appropriate metal gates. In addition, the charge in the semiconductor is not exactly a sheet charge, but exhibits some finite thickness. Even in the classical model of the surface charge the conductive channel has some finite width. An in-depth analysis of the conductive channel must consider quantum mechanical confinement effects of the carriers within a surface potential well and this results in an additional thickness of the conductive channel [5–7]. The net result is that, the charge centroid of the conductive channel is some distance below the dielectric–semiconductor interface. This adds an additional thickness to the oxide, which must be included in an expression describing the charge–voltage relationship for the MOS gate. This additional thickness is estimated as 0.4 nm in the ITRS tables. This additional 0.4-nm value accounts for the difference in the EOT and inversion CET values of Figure 1.8.

In terms of the gate capacitance characteristic, this effect reduces the maximum capacitance values in either accumulation or inversion as illustrated in Figure 1.9. This shows several theoretical *C–V* curves for a 100 nm × 100 nm capacitor on a *p*-type substrate and for a 1-nm thick oxide. Several curves are shown for a substrate doping density of $4 \times 10^{18}/\text{cm}^3$ (This value is used in order to give a threshold voltage of a few tenths of a volt). Curve (a) is the ideal *C–V* curve for a 1-nm oxide assuming an ideal gate contact, such as a metal and ignoring quantum confinement channel effects. Curve (b) is for a metal gate and including quantum channel effects. Finally, curves (c) and (d) are for n^+ polysilicon gates doped to $2 \times 10^{20}/\text{cm}^3$ and $1 \times 10^{20}/\text{cm}^3$, respectively and including quantum effects. Comparing curves (a) and (b), one sees that the major influence of the channel quantum effects is to increase the voltage for channel inversion (or threshold voltage) and to lower the peak capacitance in both accumulation and inversion. As previously discussed and reported [8,9], these are due to the shift of the channel charge centroid away from the surface due to quantum confinement effects. These make the capacitor structure appear to have a larger effective thickness and lower capacitance in the accumulation region where the conductive channel is formed. This increased thickness is represented by the t_{civ} values in Figure 1.8 and as previously stated, this increase is estimated in the ITRS as about 0.4-nm independent of the physical oxide thickness. This accounts for the difference in the t_{ox} and t_{civ} curves in Figure 1.8.

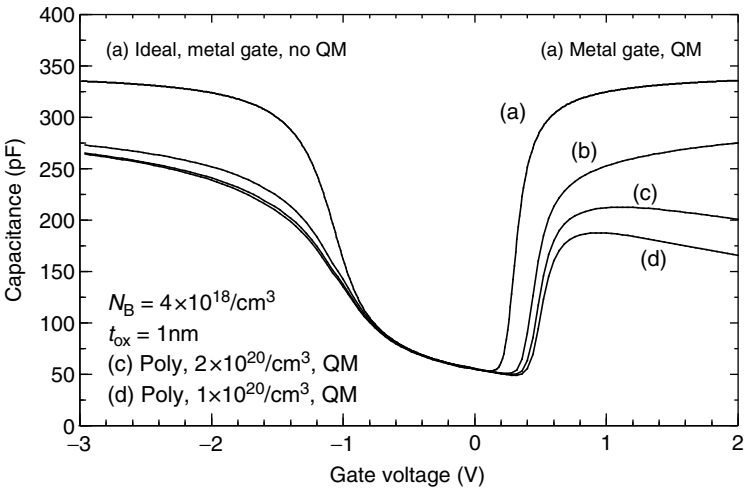


FIGURE 1.9 Theoretical *C–V* for a thin gate dielectric, including QM effects and poly depletion effects.

Curves (c) and (d) in Figure 1.9 illustrate the further reduction in inversion layer capacitance due to the use of a polysilicon gate. In the inversion region (large positive gate voltages), the polysilicon has a finite depletion layer, which adds an additional effective thickness to that of the oxide and quantum confinement effects. The capacitance values represent induced charge per unit voltage, so the differences between curves (b) and (c) or (d) represent reductions in transconductance that would occur due to polysilicon depletion. If the polysilicon doping density could be increased without limit, the polysilicon depletion effect could be minimized. However, from literature reports, it appears that it will be very difficult to get electronically active doping densities much above $10^{20}/cm^3$ for n^+ polysilicon and above the mid to upper $10^{19}/cm^3$ for p^+ polysilicon. For these doping densities, the polysilicon depletion represents a significant reduction in gate capacitance for a thin EOT dielectric.

In terms of MOS device performance, the important gate thickness related parameters in Figure 1.8 are the supply voltage, which represents maximum gate-to-source voltage and the inversion capacitance equivalent thickness, t_{civ}. A first order approximation for the C_{ox} parameter in Equation 1.3 is:

$$C_{ox} = \varepsilon_{ox}/t_{civ} = \text{Capacitance/unit area} \tag{1.5}$$

This capacitance multiplied by $V_G - V_T$ will then give the maximum channel induced charge per unit area. The possibility of using advanced dielectrics, such as high-k dielectrics in place of silicon dioxide is discussed in a subsequent section.

One important conclusion that can be gleamed from Figure 1.8 is that the projected scaling relationships have changed significantly from the 1997 to the 2004 ITRS document. The 1997 projections were essentially the case of "constant field" scaling, where all device dimensions and the voltage were projected to scale with the same factor as the technology node and the physical gate length. This type of device scaling had (in 1997) been practiced for many technology generations and can be said to constitute the "golden era of scaling" as identified in the figure for technology nodes above the 90-nm node. During this golden era of scaling, it was relatively easy (in retrospect) to scale MOS devices with relatively small changes in the device structure from generation to generation. The relative dimensions of MOS devices were close to those shown with the device cross-section illustrated in Figure 1.7. However, as Figure 1.8 shows scaling projections beyond the 90-nm node represent a significant departure from that of the golden era of scaling and away from that of constant field scaling. As subsequently discussed, this is due to the fact that several fundamental semiconductor and device limits are being approached and scaling beyond the 90-nm node becomes considerably more difficult than during the golden era of scaling.

Figure 1.10 shows the projected scaling of two other MOS device dimensions, the gate contact thickness and the spacer thickness. Overall, these are seen to scale with approximately the same scaling factor as the technology node as projected by both the 1997 and 2004 ITRS documents.

1.3.2 Performance of Scaled Devices

An obvious advantage of device scaling is the ability to pack more and more devices and electronic functionality within the same given silicon chip area. This has been the major driving force to propel the electronics industry along Moore's curve. Along with the increased packing density is of course, a desire to continue to improve the performance of the resulting circuits in terms of operating speed as well as maintaining some control over the power dissipation. Thus, it is important to understand how device scaling and in particular, how the ITRS scaling scenario affects fundamental MOS device performance. For digital circuits, three of the most important device parameters are:

a. Maximum saturated drain current (I_{dsat}) when $V_G = V_D = V_{DD}$.

b. Off-state drain leakage current (I_{off}) when $V_G = 0$, $V_D = V_{DD}$.

c. Device capacitances.

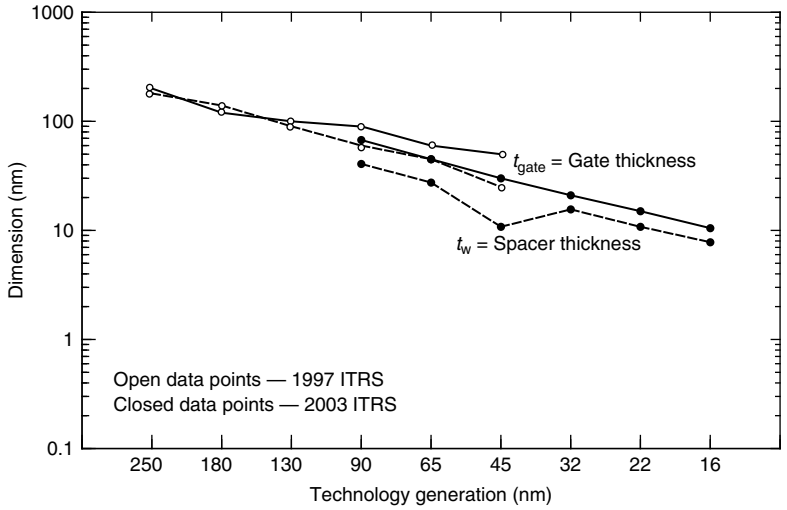


FIGURE 1.10 Projected scaling of gate and spacer thickness.

In turn, the saturated drain current depends on further device parameters, such as threshold voltage and DIBL, which then become important device parameters.

A frequently used first order model for saturated drain current is [1]

$$I_{dsat} = W v_{sat} C_{ox}(V_G - V_T - V_{dsat})$$
$$I_{dsat} = \frac{W \mu_n C_{ox}}{L_{eff}(1 + V_{dsat}\mu_n / v_{sat} L_{eff})}(V_G - V_T - \frac{1}{2} V_{dsat}) V_{dsat} \tag{1.6}$$

where μ_n is the average channel mobility and v_{sat} is the high field saturated drift velocity of the channel carriers (electrons or holes). The first equation comes from looking at the saturated velocity equation for current near the drain, while the second equation comes from integrating the channel potential equation along the channel. While these are first order model equations, they do include much of the major device physics of transport along the surface channel. Both of these equations must be satisfied at the drain saturation voltage from which we can evaluate

$$V_{dsat} = \frac{L_{eff} v_{sat}}{\mu_n}\left[\sqrt{1 + \frac{2(V_G - V_T)\mu_n}{L_{eff} v_{sat}}} - 1\right] \tag{1.7}$$

For maximum drive current, the gate voltage is set to the power supply voltage. Using this expression in either of Equation 1.6 can then give the maximum saturated drain current.

If we define the following quantities

$$Y = \frac{(V_G - V_T)\mu_n}{L_{eff} v_{sat}}, \qquad X = \frac{V_{dsat}}{(V_G - V_T)} \tag{1.8}$$

then the saturated drain current can be expressed as

$$I_{dsat} = W v_{sat} C_{ox}(V_G - V_T) F_I(X), \quad \text{where } F_I(X) = 1 - X \text{ and } X = \frac{1}{Y}\left[\sqrt{1 + 2Y} - 1\right] \tag{1.9}$$

An upper limit on $F_I(X)$ is $F_I(X) \rightarrow 1$ as carrier velocity saturation becomes the dominant factor limiting the saturation current.

The surface-channel mobility μ_n is not exactly a constant as assumed in deriving this set of device equations, but is known to depend on the effective surface electric field [10], which can be expressed as

$$E_{eff} \simeq \frac{C_{ox}}{2\varepsilon_{Si}}\left[V_G + V_T + 2\left(\frac{V_{gap}}{2} - \phi_B\right)\right] \tag{1.10}$$

where ϕ_B has the usual meaning as in Equation 1.2 and V_{gap} is a voltage corresponding to the silicon bandgap (1.1 V). Under most conditions $(V_{gap}/2 - \phi_B)$ can be neglected and

$$E_{eff} \simeq \frac{C_{ox}}{2\varepsilon_{Si}}(V_G + V_T) = \frac{\varepsilon_{ox}}{2\varepsilon_{Si}t_{civ}}(V_G + V_T) \simeq (V_G + V_T)/6t_{civ} \tag{1.11}$$

The last expression has previously been used by Hu [11] as an approximation to the effective surface field.

From these first order device current equations, there are three "field" terms that are of importance: (a) $(V_G - V_T)/L_{eff}$, (b) $(V_G + V_T)/6t_{civ}$, and (c) $(V_G - V_T)/t_{civ}$ The first field is a lateral electric field along the channel, while the second term is the effective field in the semiconductor at the surface, and the last term is the excess electric field across the gate dielectric, which is responsible for the channel charge. This is somewhat less than the oxide electric field, which can be some 20%–25% larger due to the depletion layer charge needed before establishing the conductive channel.

In terms of the MOS device performance, a critical parameter is the effective channel length, L_{eff}. However, this is not a parameter specified by the ITRS projected scaling values. This value is critically dependent on the contacting junction formation techniques and not just on the physical line width. In order to make good ohmic contact to the channel, the effective channel length must be somewhat less than the physical gate length, so that the contacting layer extends under the gate oxide for some small distance. In order to carry forward with the analysis here, some value is needed for the effective channel length. For use here, this has been assumed to be 0.7 times the physical channel length. This essentially means that the effective channel length is assumed to be one technology generation ahead of the physical gate length and to remain a fixed percentage of the physical gate length for each technology generation. Manufacturing problems associated with achieving all the ITRS parameters are covered in subsequent sections.

Figure 1.11 shows plots of these three fields according to the ITRS scaling rules for each major technology node. The open data points are obtained from the 1997 roadmap values, while the open data points are from the projected scaling in the 2004 roadmap. First, we note that the three fields are approximately constant using the 1997 parameters. This is again indicative of the essentially constant field scaling during the golden age of scaling. For the 2004 scaling projections, one sees that the vertical fields, E_c and E_{eff} in the figure are approximately constant for projected nodes of 65 nm and below. There is a projected increase in these fields by about a factor of 2 in the 2004 ITRS values as compared with the 1997 ITRS values. This increase will have two effects: (1) a possible increase in gate leakage due to the increased gate field and (2) a decrease in average surface carrier mobility due to the increased effective surface field.

The major change reflected in the 2004 projections is the greatly changed functional nature of the lateral MOS device field, E_{lat} in Figure 1.11. This is due to the reduced scaling of supply voltage relative to the scaling of the effective channel length. In terms of MOS device operation, this will enhance velocity saturation effects as the point along the channel where velocity saturation occurs will occur closer to the source contact with the increased average lateral field. In terms of the device current of Equation 1.9, this will make the $F_I(X)$ factor closer to unity. In summary, for the device electric fields, the 2004 projected scaling is essentially a constant field scaling for the vertical MOSFET fields and an ever increasing lateral device average field.

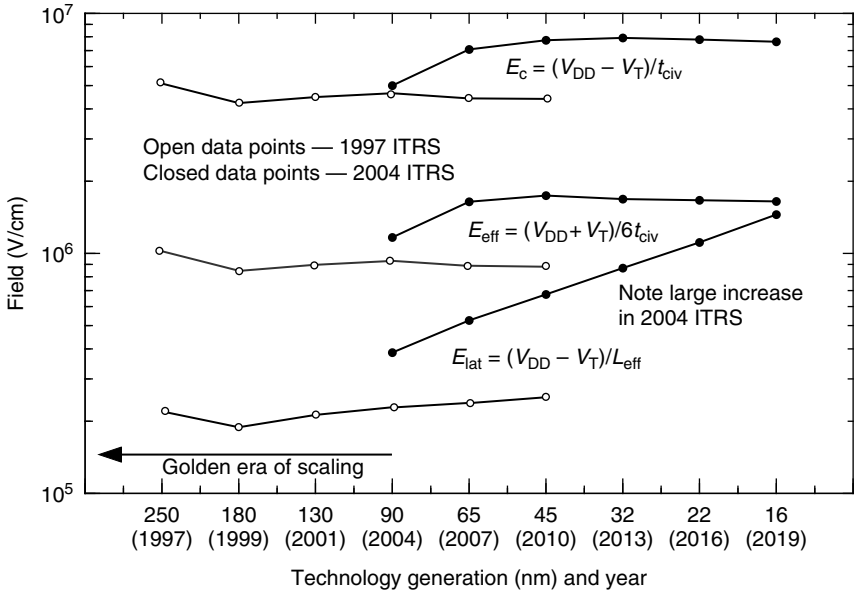


FIGURE 1.11 Projected scaling of various device field parameters.

To complete the model parameters, a value of mobility and saturation velocity is needed. For the carrier saturation velocity, the value of 1×10^7 is frequently used for long-channel devices. There may be some tendency for this value to increase at small device dimensions because of velocity overshoot effects. However, for an initial calculation this will not be considered. The electron mobility is strongly dependent on the effective surface electric field because of surface scatterings. At a peak effective surface field of 1×10^5 V/cm, the surface mobility of electrons varies from approximately 250 cm^2/V s to approximately 320 cm^2/V s [10]. So for, the 1997 scaling an average value of 280 cm^2/V s can be used. For the 2004 scaling projections, the peak surface effective field is expected to increase to about 2×10^5 V/cm. Unfortunately, this increase in effective surface field will result in a reduction in surface mobility by about a factor of 2 (in this region of field, the surface mobility is approximately inversely proportional to the surface field). In order to illustrate the separate effects of the lateral field and the reduced surface mobility, results will be shown for the 2004 parameters using both 280 and 140 cm^2/V s.

The consequences of these projected scaling rules can be seen with respect to the first order device parameters in Figure 1.12. For the 1997 constant field scaling, the parameters are essentially constant for each technology generation as expected. For the 2004 scaling, the drain saturation voltage is seen to be an ever decreasing function of the technology generation, being about 0.33 times the excess voltage above threshold for the 90-nm node and decreasing to about 0.20 times the excess voltage for the 16-nm node. This is a direct consequence of the increased lateral field as seen in Figure 1.11. This in turn causes an increase in the F_I device equation factor as seen in the figure. This has a simple physical interpretation of representing the fraction of the V_{sat} velocity that the carriers have achieved when current saturation occurs. Because of the higher lateral fields, the figure shows that this factor slowly increases with the shrinking of device dimensions.

The final important parameter is the device current or rather the projected current per unit length of gate. This is shown for the first order model and the above parameters in Figure 1.13. For the 1997 constant field parameters, this value is constant at about 760 μA/μm (or mA/mm using the ITRS notation). For the 2004 projected scaling parameters, the maximum saturated drive current is projected to increase somewhat from the 90-nm node to the 45-nm node, but then stay relatively constant to the 16-nm node. The model values are close to the ITRS desired roadmap values to the 45-nm node, but then

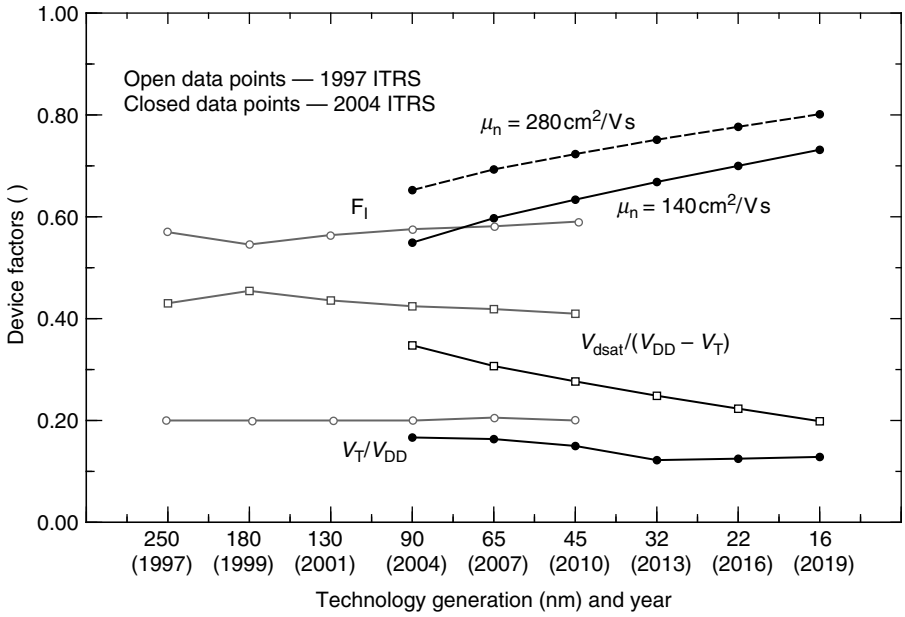


FIGURE 1.12 Projected scaling of various metal oxide semiconductor device current density factors.

fall below the ITRS projected values. By comparing the saturated current values to the model equations, one can see that most of the projected increase in drive current comes about from the increase in the $C_{ox}(V_G - V_T)$ value or in the increase in the vertical electric field seen in the effective and oxide fields in Figure 1.11. A small amount of the increase comes from the increased lateral field because of the upper limit of the carrier saturation velocity.

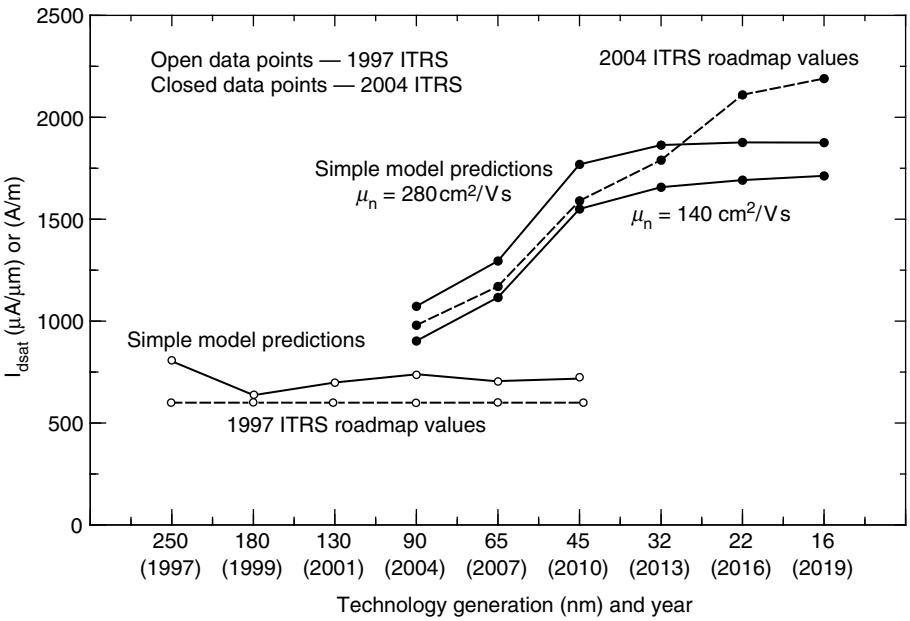


FIGURE 1.13 Projected scaling of *n*-metal oxide semiconductor saturated drain current at maximum gate voltage.

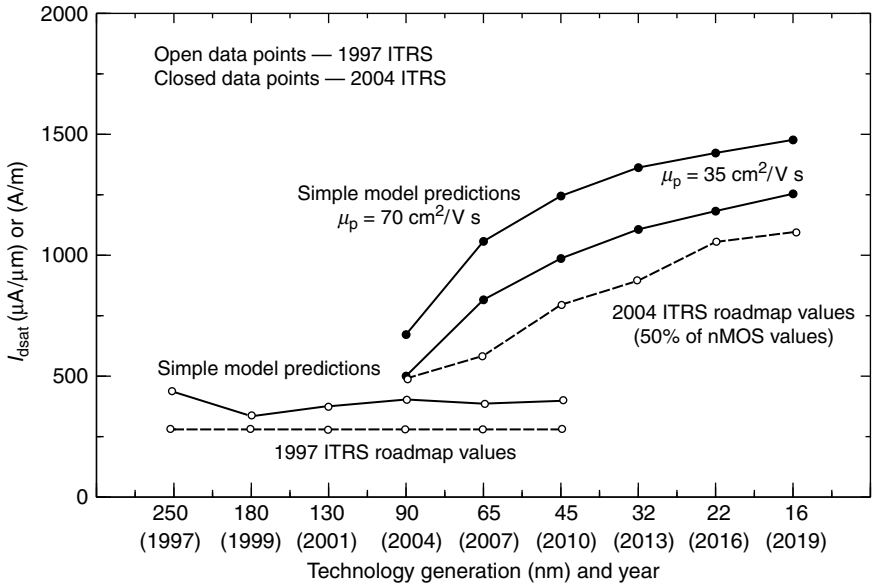


FIGURE 1.14 Projected scaling of *p*-metal oxide semiconductor saturated drain current at maximum gate voltage.

A similar set of calculations can be made for holes, using a surface mobility of approximately 70 and 45 cm^2/V s for the 1997 and 2004 parameters, respectively, and using a v_{sat} value of 8.4×10^6 cm/s. These values give the projected *p*-channel current values shown in Figure 1.14. In this case, one can again see that there is a considerable increase in the projected saturation current using the 2004 scaling parameters. In this case, the increase comes both from the increased lateral field and from the projected increase in the surface charge density due to the increased surface field with the larger power supply voltage. In this case, the simple model predicts *p*-channel saturated current values slightly larger than the ITRS projected values. The relative enhancement of the *p*-channel current with respect to the *n*-channel current arises in the simple model, because the ratio of saturated velocities for holes-to-electrons is larger than the corresponding ratio of low field mobilities. The increased lateral field, with scaling, pushes both types of devices closer to the velocity saturation limit, hence increasing the *p*-channel saturation current by a larger factor than the *n*-channel saturation current. The relative ratio of hole-to-electron current remains essentially constant for constant field scaling in both the vertical and lateral dimensions.

The simple model neglects at least two important factors that must be considered in any more realistic estimate of current drive. First, the DIBL and channel-length modulation effects cause an increase in current above that predicted by such a first order model. As indicated in Figure 1.5, this can cause a relatively large percentage increase in the available drive current at the supply voltage value. A reasonable enhancement value might be a 25%–30% increase in current. A second neglected factor is the effects of source and drain series resistance. These resistances reduce the current drive by effectively reducing the internally applied gate-to-source voltage. In order to minimize this latter effect, the source/drain contacts must be very carefully constructed and this becomes more difficult as device dimensions shrink. Typical source/drain resistances can easily reduce the available current drive by 25%–30%. Thus, these two neglected factors tend to somewhat offset each other with one increasing current drive and the other decreasing current drive. For the purpose here, it will be simply assumed that these effects tend to offset each other and that the simple model gives reasonably good first order approximations to the available current drive at the projected ITRS scaling. In any case, the user should not place high confidence in the exact projected values, but only accept them to within some 25%–30% of accuracy. Difficulties in

achieving the scaled parameters and possible enhancements are discussed in a subsequent section on manufacturing issues.

Device current drive is important because it relates to other important parameters, such as switching speed or logic gate delay, and power dissipation. In terms of logic gate delay, the important parameters can be obtained from the simple equation

$$I_{\text{dsat}} = C_{\text{L}}\Delta V_{\text{L}}/\Delta t \tag{1.12}$$

where C_{L} is the gate load capacitance and ΔV_{L} is the change in load voltage in time Δt. From this, we can calculate a gate delay as

$$\tau_{\text{d}} = C_{\text{L}}\Delta V_{\text{L}}/I_{\text{dsat}} \tag{1.13}$$

This illustrates the importance in saturated drive current in achieving ever smaller gate delays, which can translate into higher system clock frequencies. While the load capacitance is a complex combination of device capacitances and wiring layout capacitances, it is generally agreed that to the first order, the load capacitance is expected to vary directly with the lithography dimension or with the technology generation. In the golden era of scaling, the saturation current per unit gate length was approximately constant, implying that the saturation current was also varying directly with the technology generation. Also the logic voltage level was scaling directly with the technology generation. Thus, if we let $\alpha(\simeq\sqrt{2})$ be a scale factor (or perhaps the inverse scale factor) per technology generation, then for constant field scaling

$$C_{\text{L}} \propto 1/\alpha, \qquad \Delta V_{\text{L}} \propto 1/\alpha, \qquad I_{\text{dsat}} \propto 1/\alpha \quad \text{and} \quad \tau_{\text{d}} \propto (1/\alpha)(1/\alpha)/(1/\alpha) = 1/\alpha \tag{1.14}$$

The last line is the most important, where the constant field scaling gives a gate delay, which decreases by approximately 0.707 $(1/\alpha)$ per technology generation. This affords a path for the system level clock frequency to increase by about 40% (by a factor $\alpha = 1.414$) for each technology generation. This has been used by the industry during the golden era of scaling to not only increase the packing density and functionality, but also to rapidly increase the clock frequency of microprocessors by many factors of 2.

The projected scaling in the 2004 ITRS makes it considerably more difficult to continue decreasing the gate delay. This arises because the highly desired decrease in logic voltage level has been considerably slowed. A detailed look at the projected scaling of the power supply voltage in Figure 1.8 shows that the decrease has changed to about 15% per technology generation rather than the previous 30% per generation. To a first order approximation then

$$\Delta V_{\text{L}} \propto 1/\sqrt{\alpha} \tag{1.15}$$

In order to maintain a gate delay that then improves in the same manner as for constant field scaling, one must require a slower decrease in saturation drain current than with the technology generation. The required decrease in saturation current is then the same as that of the power supply voltage. If we then use the scaling factors:

$$C_{\text{L}} \propto 1/\alpha, \qquad \Delta V_{\text{L}} \propto 1/\sqrt{\alpha}, \qquad I_{\text{dsat}} \propto 1/\sqrt{\alpha} \tag{1.16}$$

then one finds

$$\tau_{\text{d}} \propto 1/\alpha \quad \text{but it requires that } I_{\text{dsat}}/W \propto \sqrt{\alpha} \tag{1.17}$$

This last equation summarizes the important gate delay projections from the 2004 ITRS. The gate delay is projected to continue to decrease directly with the technology generation, but in order to achieve this, the saturated current per gate width must be forced to increase with each technology generation.

The 2004 ITRS clearly states that the proposed scaling scenario is based upon this important assumption. The projected linear increase in saturated current per unit width seen in Figure 1.13 and Figure 1.14 for *n*- and *p*-channel devices is a direct consequence of this assumption. Under this scenario, four technology generations require a doubling of the saturated drive current per unit length and this is essentially the projection seen in Figure 1.13 and Figure 1.14 for the ITRS projected values.

Another possible scenario would be to accept a device scaling that produces a constant current per unit width and accept a slower decrease in gate delay per technology generation. If one considers keeping the saturated current per unit gate length constant then this leads to a $\tau_d \propto 1/\sqrt{\alpha}$ instead of the relationship of Equation 1.17. In this case, four technology generations would be required to double the clock frequency instead of the two generations with the projected ITRS scaling.

Another very important scaling parameter is the power dissipation per gate and the power density or power per unit gate area. Traditionally, power dissipation in CMOS circuits has been dominated by the dynamic switching power because of the very low gate and drain currents in the two stable logic states. Considering only this dynamic power for one gate, this can be expressed as

$$\begin{aligned} p_g &= \frac{1}{2} C_L V_{DD}^2 f \\ \frac{p_g}{A_g} &= \frac{C_L V_{DD}^2 f}{2A_g} \end{aligned} \tag{1.18}$$

where C_L is the load capacitance that is being switched at some frequency f. The first form expresses power per gate, while the second form represents power density or power per unit gate area A_g. While both of these are important, the power density is probably the more important factor as present ICs are close to perceived limits in terms of handling power per unit area. For constant field scaling with the scaling of Equation 1.14 this becomes

$$\frac{p_g}{A_g} \propto \frac{(1/\alpha)(1/\alpha)^2}{(1/\alpha)^2} f \propto f/\alpha \tag{1.19}$$

To keep constant power per unit gate area, one can then increase the clock frequency directly with the scaling factor per generation or double the clock frequency with every two technology generations. Of course, one must also consider total chip power, which for increasing chip areas would continue to increase with each generation. However, if power per unit area is the limiting factor, the decreased gate switching speed can be directly employed with the increased system level clock frequencies.

Things have now, however, changed with the 2004 ITRS projections due to the slower rate of decrease in supply voltage. The decrease in voltage squared will no longer offset the decreasing area and in fact, both the capacitance and the voltage decrease are needed to offset the decreasing area and one obtains with the 2004 ITRS scaling

$$\frac{p_g}{A_g} \propto \frac{(1/\alpha)(1/\alpha)}{(1/\alpha)^2} f \propto f \tag{1.20}$$

The power per unit area is now directly proportional to the clock frequency and any increase in the clock frequency will result in an increased power density! This is an important implication of the slowed voltage scaling of the 2004 ITRS. IC design and performance is no longer in the golden era of scaling that prevailed for many generations before about the 90-nm node.

For system applications, one needs to take a little more general interpretation of the clock frequency in the above equations. For a large system not all gates will be operating at any given time and only the active gates contribute to the dynamic power dissipation. Thus, we can expand the interpretation of the

frequency to an equivalent clock frequency with

$$f = f_{\text{clk}} f_{\text{act}} \tag{1.21}$$

where the two factors are the clock (clk subscript) frequency and the fraction of gates that are active (act subscript). It is then the product of the system level clock frequency and the fraction of active gates that must remain constant with the 2004 ITRS scaling. Thus, the clock frequency could still be increased by 40% per technology node, provided an additional 40% of the gates were inactive at all times and the average power per gate would not change. These tradeoffs obviously offer many possible alternatives, such as the possibility of doubling the circuit functionality and keeping the same clock frequency at a new technology node. For example, two microprocessors at the same clock frequency could be implemented, instead of increasing the clock frequency and reducing the percentage of active circuits.

In addition to the dynamic switching power, static power dissipation is becoming a major design parameter for scaled circuits. The MOS off-state drain leakage current, $I_{\text{d,leak}}$, consists of components due to (1) gate leakage, (2) drain-substrate leakage, and (3) drain-to-source leakage. As covered in more detail later, all of these components tend to increase exponentially with the decreasing device dimensions. How to control these in the manufacturing process is a major challenge. For the purposes of proper circuit operation and power management, these parasitic and undesirable currents must be kept below some upper limit. Static off-state power dissipation can be expressed as

$$p_{\text{off}} = V_{\text{DD}} I_{\text{d,leak}} \tag{1.22}$$

Limits for the allowable leakage current comes not from the fundamental device limits, but from circuit and system limits on the allowable values of static off power that can be managed in a given IC application. In general, one desires a low value of off-to-on state leakage current. Acceptable past values for MOS devices have been in the range of 1×10^{-5} to 1×10^{-4} and acceptable values were not even specified in the 1997 ITRS. Figure 1.15 shows the off-to-on current ratios specified in the 2004 ITRS for the 90- to 22-nm technology nodes. Three curves are shown for (1) HP, (2) low operating power (LOP), and (3) low standby power (LSP). While there is some increase projected in the allowable leakage with scaled devices, we can see that the projected leakage is in the range of 1×10^{-4}, 1×10^{-5} and 1×10^{-7} for

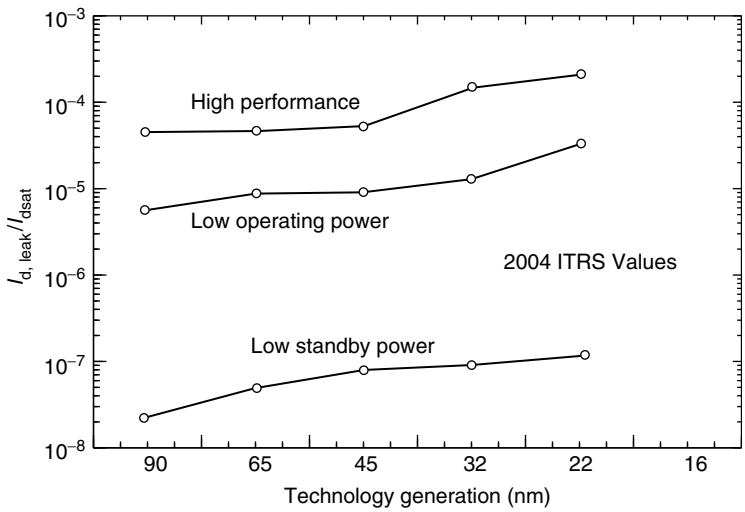


FIGURE 1.15 Projected ratio of allowable offstate leakage current to saturated drain current.

the three applications, respectively. Difficulties in achieving these numbers are further discussed in the manufacturing section. Static power dissipation at the system level is becoming much more of a problem with further scaling of devices and is a major consideration with highly desirable low power applications and especially LSP applications.

In terms of the future scaling of off-state power per unit device area, consider the best case scenario, where leakage current remains a constant multiple of saturated device current. In this case and for the 2004 ITRS scaling factors of Equation 1.16, one obtains

$$\frac{p_{\text{off}}}{A_{\text{g}}} \propto \frac{(1/\sqrt{\alpha})(1/\sqrt{\alpha})}{(1/\alpha)^2} = \alpha \tag{1.23}$$

This indicates that the expected off-state power per unit gate area is expected to increase with future generations even with a fixed off-to-on current ratio and will increase even faster with the increased off-state current shown in Figure 1.15. This is another indication that power management is to become more and more of a major technology limiter. Various power-down techniques can perhaps be used to help and manage this problem at the circuit and system level, as there does not appear to be any solution at the basic device level.

These various scaling rules can be conveniently summarized in a table, such as Table 1.1. For the most generalized scaling, it is assumed that different scaling could possibly be used with the following parameters: (1) device width and length (α_{L}), (2) oxide thickness (α_{ox}), (3) supply voltage (α_{v}), (4) saturation current (α_{I}), (5) wiring capacitance (α_{W}), and (6) effective clock frequency (α_{F}). In the case of traditional constant field scaling, all of these parameters have the same value. For the major technology nodes, the length scaling factor is approximately 1.414. The third column in the table summarizes the approximate scaling for the 2004 ITRS, where the supply voltage and oxide thickness scales at approximately half the rate of the technology generation. This table is a convenient summary of the ITRS scaling discussed previously in this section, and summarizes some of the difficulties this presents in regard to control the power density and increase the clock frequency for future ICs.

TABLE 1.1 Technology Scaling Rules for Generalized Scaling, Constant Field Scaling and 2004 ITRS Scaling Rules

Physical Parameter	Generalized Scaling	Constant Field Scaling (Golden Era or Scaling)	2004 ITRS Scaling ($\alpha_{\text{L}}=1.414$ for Each Major Technology Node)
Device dimensions (L,W)	$1/\alpha_{\text{L}}$	$1/\alpha$	$1/\alpha_{\text{L}}$
Metal oxide semiconductor (MOS) oxide capacitance thickness	$1/\alpha_{\text{ox}}$	$1/\alpha$	$1/\sqrt{\alpha_{\text{L}}}$
Supply voltage	$1/\alpha_{\text{v}}$	$1/\alpha$	$1/\sqrt{\alpha_{\text{L}}}$
I_{dsat}	$1/\alpha_{\text{I}}$	$1/\alpha$	$1/\sqrt{\alpha_{\text{L}}}$
I_{dsat}/W	1	1	$\sqrt{\alpha_{\text{L}}}$
Doping density	$\alpha_{\text{L}}^2/\alpha_{\text{v}} \leftrightarrow \alpha_{\text{L}}^2$	$\alpha \leftrightarrow \alpha^2$	$\alpha_{\text{L}}^{3/2} \leftrightarrow \alpha_{\text{L}}^2$
Wiring length or wiring cap	$1/\alpha_{\text{W}}$	$1/\alpha$	$1/\alpha_{\text{L}}$
MOS device area	$1/\alpha_{\text{L}}^2$	$1/\alpha^2$	$1/\alpha_{\text{L}}^2$
Logic gate area	$1/\alpha_{\text{W}}^2$	$1/\alpha^2$	$1/\alpha_{\text{L}}^2$
Wiring capacitance	$1/\alpha_{\text{W}}$	$1/\alpha$	$1/\alpha_{\text{L}}$
Logic gate delay (τ_{d})	$\alpha_{\text{I}}/(\alpha_{\text{W}}\alpha_{\text{v}})$	$1/\alpha$	$1/\alpha_{\text{L}}$
Logic clock frequency	α_{F}	α	α_{F}
Gate power dissipation	$\alpha_{\text{F}}/(\alpha_{\text{W}}\alpha_{\text{v}}^2)$	$1/\alpha^2$	$\alpha_{\text{F}}/\alpha_{\text{L}}^2$
Gate power density	$\left(\alpha_{\text{F}}\alpha_{\text{L}}^2\right)/\left(\alpha_{\text{W}}\alpha_{\text{v}}^2\right)$	1	α_{F}
Off state current	$1/\alpha_{\text{I}}$	$1/\alpha$	$\geq 1/\sqrt{\alpha_{\text{L}}}$
Off state power density	$\alpha_{\text{L}}^2/(\alpha_{\text{v}}\alpha_{\text{I}})1$	1	$\geq \alpha_{\text{L}}$

1.4 Manufacturing Issues and Challenges

The scaling projections in the previous section and those developed by the industry in the ITRS represent a desirable set of device goals in terms of device dimensions and performance that are needed to continue the exponential growth in IC system performance in future years. The scaling projections and device performance projections do not necessarily take into account physical material and device structure limits that may make it very difficult, if not impossible to achieve these goals. Some considerations of physical limits are incorporated into the scaling projections, especially in regard to limits on gate dielectric thickness. These considerations are the fundamental reason that the projected scaling for power supply voltage and oxide thickness were fundamentally changed between the 1997 and 2004 ITRS document. It was simply realized that achieving the projected oxide thicknesses in the 1997 document were physically impossible. Likewise, some of the projections in the 2004 ITRS for device performance and/or size scaling may prove to be physically impossible to achieve. This section presents some of the major challenges faced by the semiconductor industry in achieving the dimensional and performance scaling of CMOS devices discussed in previous sections. Very serious manufacturing problems exist in achieving the projected scaling to the end of conventional CMOS devices.

1.4.1 MOSFET Gate Stack Issues

The gate stack design is the key part of the MOSFET. It has been recognized for some time that continuing to scale the gate stack with the same materials of silicon dioxide and polysilicon will not produce acceptable devices for gate oxide thicknesses approaching 1 nm and beyond. Unlike in the golden era of scaling, innovative approaches, in both the gate dielectric and contacting material, are necessary. The gate stack (see Figure 1.7) is composed of (1) the oxide–silicon interface, (2) the gate insulator, and (3) the gate contact layer. For devices approaching the end of the roadmap, the gate oxide performance at a projected thickness of less than 1 nm is critical. The SiO_2 cannot be utilized at these dimensions. Aside from reproducibility and manufacturability issues, tunneling currents are the major problem. Figure1.16 shows experimental gate oxide tunneling currents for oxide thickness ranging from 3.5 to 1.4 nm at an applied voltage, where direct tunneling dominates the current [12]. While these currents are large, they must be evaluated in terms of what might be an acceptable level in a particular device application. It has been experimentally shown that good device characteristics can be achieved with very thin oxides, provided the gate length is sufficiently short [13]. A gate current of 1 A/cm^2 was suggested early on as an upper limit [12]. However, it is now generally accepted that the allowable gate current can be much larger from a device performance point of view. The ITRS document for HP devices lists acceptable gate current densities as high as 1×10^4 A/cm^2. It is important to understand how such large gate currents can be considered as acceptable. If one assumes that an I_{off}/I_{on} ratio of 10^{-4} as shown in Figure 1.15 for HP devices is acceptable and that half the off-state leakage can be due to gate current, then we can write

$$\frac{I_g}{A_g} = \frac{I_g}{WL_p} \leq 5\times10^{-5}\frac{I_d}{WL_p} \tag{1.24}$$

The worst case (lowest limit) will occur for the minimum drain current, which occurs for *p*-MOS devices and ranges in terms of current per unit width from about 500 to 1000 A/m for the 90- to 16-nm nodes. Then using the lower value, we can obtain

$$\frac{I_g}{A_g} \leq \frac{(2.5\times10^{-4}\text{A/cm})}{L_p} = \frac{(2500\text{A/cm}^2)}{(L_p/\text{nm})} \tag{1.25}$$

For the 90- to 16-nm technology generation, this gives values from about 68 to 417 A/cm^2. For an *n*-MOS gate, the limits would be about a factor of 4 larger, keeping the same ratio of off-to-on current. As from the previous section, the 2004 ITRS projects even larger ratios of off-to-on current for the smaller device dimensions. In any case for the HP devices, the gate current density can be quite large and cannot degrade the device performance. For LOP and LSP device applications, the current density must be considerably lower. Even though the current densities may seem high from conventional wisdom about a dielectric, the total currents can be a small fraction of the saturated drain current. Limits to gate current come not from degradation in device performance, but from overall chip power considerations.

It is now certainly known that SiO_2 for the MOS gate will be used at considerably smaller device dimensions than thought some few years ago. However, the data in Figure 1.16 illustrates the gate leakage problem with SiO_2. For a given voltage, the current increases about a factor of 10 for each decrease of about 2-nm in oxide thickness due to direct tunneling through the oxide. The 2004 ITRS projects that oxynitrides, which are somewhat better than pure SiO_2, will not satisfy the leakage current requirements at the 45-nm node and beyond, and that improved gate dielectric materials are needed for future technology generations. The purpose here is not to try to project exactly when alternative dielectrics will be required, but to discuss the potential advantages of improved gate dielectrics.

The potential for reduced gate leakage with advanced dielectrics can be seen from the first order equation for carrier tunneling probability T through a barrier of height V_b and thickness t_d [14]:

$$\begin{aligned} &\text{(a) Low Fields :} \quad T \propto \exp\left(-\sqrt{\frac{2mqV_b}{\hbar^2}}t_d\right) \\ &\text{(b) High Fields :} \quad T \propto \exp\left(-\frac{4}{3}\sqrt{\frac{2m}{\hbar^2}}\frac{(qV_b)^{3/2}}{qV}t_d\right) \end{aligned} \tag{1.26}$$

where V/t_b is the electric field across the barrier. From this, we can identify the important material and device parameters determining tunneling current:

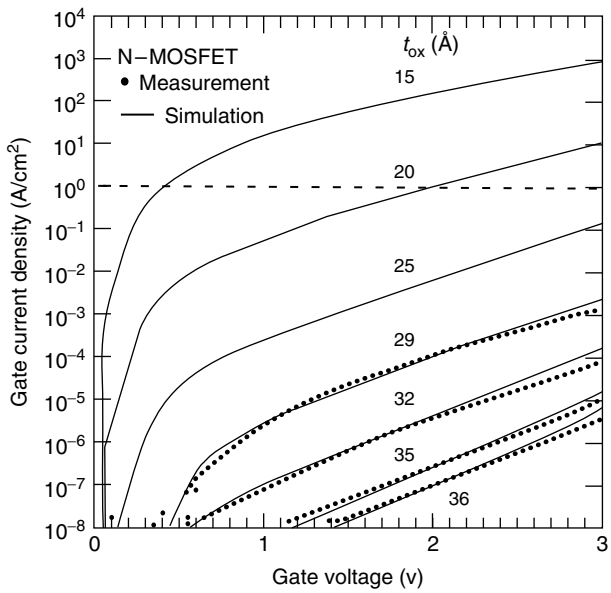


FIGURE 1.16 Measured and simulated gate currents for thin SiO_2 gate stacks. (From Taur, Y., D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S-H. Lo, G. A. Sai-Halong et al., *Proc. IEEE*, 85, (1977): 486.)

$$\begin{aligned} &\text{(a) Low fields :} \quad \sqrt{mV_b}t_d \\ &\text{(b) High fields :} \quad \sqrt{m}V_b^{3/2}t_d \end{aligned} \tag{1.27}$$

When considering different possible gate dielectric materials, a device constraint is the need to maintain the same capacitance across the dielectric, so that the same channel charge can be induced into the semiconductor with the same gate voltage. From the capacitance equation, we can see that for the same capacitance the dielectric thickness would vary directly with the dielectric constant as:

$$t_d = \frac{\varepsilon_d}{C} \propto \varepsilon_d \tag{1.28}$$

We can thus define a tunneling current "figure of merit" which captures the important material and device parameters for comparing the tunneling current of different dielectrics as:

$$\begin{aligned} &\text{(a) Low fields :} \quad \sqrt{mV_b}\varepsilon_d \\ &\text{(b) High fields :} \quad \sqrt{m}V_b^{3/2}\varepsilon_d \end{aligned} \tag{1.29}$$

Since these parameters appear as negative quantities in the exponential factor, whichever material has the largest value of this parameter will, in theory, have the lowest tunneling current. If only the dielectric constant varied between materials, one would always want a dielectric material with a very high dielectric constant. However, there is a strong tendency for materials with high dielectric constants to also have smaller bandgaps, which must translate into lower barrier heights. Figure 1.17 shows some of the most important potential high-*k* dielectrics and their bandgap values vs. dielectric constant. While the high-*k* materials tend to have smaller bandgaps from the above figure of merit, one can see that a doubling of the dielectric constant is more important than a reduction in the barrier height by a factor of 2. The tunneling current is very sensitive to the tunneling barrier thickness.

Table 1.2 shows values of these figure of merit parameters for several materials, assuming that the effective mass is the same for all the materials. This is done because of the lack of reliable effective mass values for the various materials. Two sets of values are shown for several of the materials because of some uncertainty with regard to dielectric constant and barrier height. When two values are given, the first set corresponds to a very optimistic set and the second term corresponds to the low range of

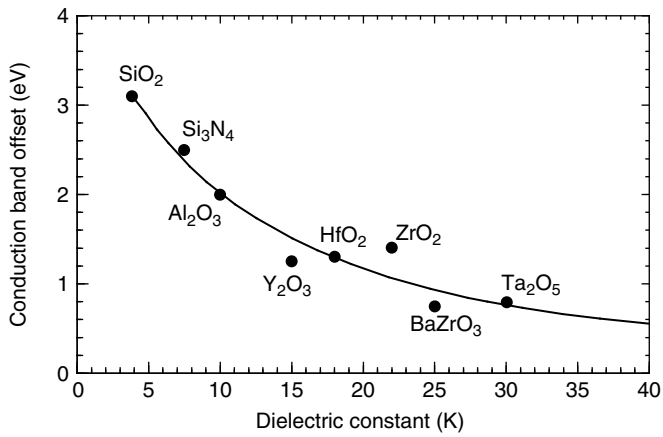


FIGURE 1.17 Variation of conduction band offset with Si for various high-*k* dielectrics.

TABLE 1.2 Tunneling Figure of Merit Parameters for Several Dielectrics

Material	ε_d	V_b(V)	Low Fields $V_b^{1/2}\varepsilon_d$	High Fields $V_b^{3/2}\varepsilon_d$
SiO_2	3.9	3.0	6.75	20.3
Si_3N_4	7.8	2.0	11.0	22.1
Ta_2O_3	25.0	1.5	30.6	45.9
TiO_2	25.0	2.1	36.2	76.1
HfO_2	50.0	1.0	50.0	50.0
ZrO_2	30.0	1.0	30.0	30.0
	28.0	1.5	34.3	51.4
	20.0	1.4	23.7	33.1
	24.0	1.4	28.4	39.7
	20.0	1.3	22.8	29.6

expected values. In general, more improvement is seen to be theoretically possible at low fields than high fields; and higher dielectric constant materials offer larger potential improvements even though they tend to have reduced barrier heights. The larger parameters in the table would be expected to show the lowest leakage currents, all other things being equal. It must be remembered that the values in the table do not account for any possible differences in tunneling effective mass between the different materials. In most cases, one would expect the tunneling mass to be smaller in the higher-*k* dielectrics and to negate some of the potential advantage of the higher-*k* materials. The case of Si_3N_4 is interesting because the low field parameter projects an improvement in the tunneling current, while the high field parameter projects little improvement in the tunneling current over that of SiO_2. For the present and future MOS devices, the low field regime is the most important and thus, one does expect an improvement with nitrides and oxynitrides, and oxynitrides have been verified by many investigators to show reductions in tunneling currents [15,16]. This is the first route for manufacturers toward reducing leakage currents in MOS devices and has been implemented already by the IC manufacturers.

A high-*k* dielectric such as HfO_2 with a dielectric constant of 28 can be made about $28/3.9 = 7.2$ times physically thicker than SiO_2 and give the same capacitance or same inversion layer charge at the same gate voltage. In such a case, a 7.2-nm thick HfO_2 dielectric would have an EOT of only 1.0-nm and may be potentially much easier to manufacture and control. The ITRS anticipates the use of alternative high-*k* dielectrics for future technology generations and expresses the dielectric thickness in terms of EOT values. In terms of real physical thickness (t_d) of a dielectric layer, the EOT can be expressed as

$$\text{EOT} = t_d(\varepsilon_{ox}/\varepsilon_d) \tag{1.30}$$

In experimental measurements such as capacitance measurements, only the EOT value can be determined unless one has a separate independent measurement of physical thickness or dielectric constant. However, for MOS device applications it is the EOT value that is important in terms of induced channel charge in inversion.

In addition to appropriate barrier height and dielectric constant, any alternative gate dielectric must form a stable compound and stable interface with silicon as well as the gate contact material at any subsequent processing temperatures. There are potential problem areas with all potential high-*k* dielectrics and the search for the most appropriate high-*k* gate dielectric represents an important research area. The most promising material appears to be HfO_2 and alloys of this with silicon, the so called Hafinum silicates.

The quality of the dielectric-silicon interface is critical for achieving high channel mobility. It is not clear that the required low surface state densities, low fixed charge and smooth interface can be achieved with any material combination other than silicon–silicon dioxide. If such an interface layer is required, it may be extremely difficult to achieve EOT below about 0.5 nm as this may be the range of interface oxide needed for good oxide-silicon properties. Some recognition of this is inherent in the ITRS projections as the end-of-roadmap EOT value is projected to be approximately 0.5 nm.

The need for high-*k* gate dielectrics has been at the forefront of gate stack research since the 1997 ITRS and much research has been done on a wide variety of MOS devices with high-*k* dielectrics. Many researchers have demonstrated orders of magnitude improvements in the gate tunneling currents over pure SiO_2 for EOT values in the range of 1–1.5 nm [17,18]. Researchers have also reported capacitors and transistors with EOT values approaching the 0.5 nm range. The fundamental approach for reducing tunneling currents as expressed in the figure of merit of Equation 1.29 is now well established. However, some problems have continued to be prevalent in the search for the ideal high-*k* gate dielectric. Many of the lowest EOT experiments have shown unacceptable reductions in surface mobility. Also many samples have shown unacceptable field-dependent threshold voltage shifts. The ideal way to transition an MOS surface interface from silicon into a high-*k* dielectric has proven to be a somewhat elusive goal, but progress is continually being made and the future of scaled MOS devices seems to include, of necessity, a high-*k* material in order to reduce the gate tunneling currents to acceptable levels.

Gate oxide reliability has always been a major concern as MOSFET devices are scaled. Operating voltages must remain below some maximum value, which is typically set by the requirement of a 20-year lifetime for devices when operated at maximum voltages. Previously reported work on SiO_2 has shown that gate oxide lifetime is determined primarily by the electric field applied to the oxide. In the era of constant field scaling, the oxide field tended to stay constant with scaled device dimensions. To the same order of approximation as involved in Equation 1.10, one can write:

$$E_{ox} \simeq (V_{DD} + (V_{bi} - 2\phi_B)/2)/t_{civ} \simeq V_{DD}/t_{civ} \tag{1.31}$$

Since the threshold voltage tends to be about 20% of the supply voltage, the oxide equivalent dielectric field according to the ITRS projected scaling rules should be about 1.24 times the E_c curve shown in Figure 1.11. Assuming this to be the case, the oxide equivalent dielectric field should remain below about 1×10^7 V/cm. The term equivalent dielectric field is used here as the dielectric will most likely not be pure SiO_2, but some more appropriate high-*k* material. Previous studies of SiO_2 have shown that it can achieve 20-year lifetimes, provided the oxide field remains below about 8×10^6 V/cm [11]. The values in Figure 1.11 prior to the 65-nm node are consistent with this as an upper limit on the oxide field. Thus, there are potential dielectric reliability problems with scaling beyond the 90-nm node as the gate field is projected to increase.

An additional factor that is somewhat unknown is the reliability of dielectrics with the very large tunneling currents that are projected in the dielectrics. Carriers transiting the oxide by pure tunneling are not expected to interact with the atoms creating any defects. However, some small fraction of the carriers will interact in the dielectric and create defect centers, which contribute to reliability problems. If a dielectric is sufficiently thin, charge trapping should not be a problem. For SiO_2, it has previously been observed that charge trapping effects tend to become negligible at thicknesses around 4–6 nm, because any trapped carriers can rapidly tunnel out of the oxide into the gate or substrate [19,20]. Hot carrier reliability is expected to greatly improve voltages below approximately 3 V, since few carriers can gain sufficient energy to create interface states or shallow oxide traps. Finally, as the dielectric EOT becomes thinner, the amount of interface charge or dielectric charge needed to shift the device threshold voltage by a small voltage such as 1 mV keeps increasing, so the scaled devices are much more tolerant to interface and dielectric charges. This is one of the positive aspects of scaled devices.

For the high-*k* dielectrics which are expected to be used with the scaled devices, reliability is still somewhat unknown. However, one trend in the right direction is the reduced magnitude of the dielectric field with increased dielectric constant. For example, an oxide field of 1×10^7 V/cm for SiO_2, with a dielectric constant of 3.9, corresponds to a dielectric field of less than 2×10^6 V/cm for a high-*k* dielectric with a constant of 20. This provides considerable encouragement that reliability of high-*k* dielectrics will not be a major problem at the expected fields. Some preliminary data on the reliability of high-*k* dielectrics has been encouraging, and it may in fact be possible to increase the effective dielectric field as proposed in the ITRS without degrading long-term device lifetime. As this is written, there are still

considerable problems to be resolved for the gate dielectric to be used as researchers attempt to approach EOT values in the 0.5-nm range.

In addition to the gate insulator, the gate contact is a critical part of the gate stack. This has typically been polysilicon and more typically both n^+ and p^+ polysilicon for *n*- and *p*-channel devices, respectively. In the early history of MOS devices, the replacement of metal gates by polysilicon was one of the key developments leading to rapid improvements in yield of early ICs. The MOS device has, however, continued to maintain the name MOS device. The use of polysilicon has become a major problem for continued device scaling, because the finite depletion layer associated with the gate charge in the polysilicon is causing significant drops in the current drive and transconductance of MOS devices. The polysilicon depletion effect becomes more important with scaling, because the oxide field (see Figure 1.11) remains essentially constant with scaling. This means that the charge per unit area in the gate polysilicon remains constant with scaling. In turn, the voltage drop in the polysilicon for the same gate doping density remains constant. Thus, since voltage levels decrease with scaling, the polysilicon voltage drop becomes a larger fraction of available device voltage and current drive is reduced with each technology generation. To maintain a constant dielectric field with polysilicon depletion requires a further decrease in an already very thin dielectric.

Simulated *C–V* curves for n^+ polysilicon on a *p*-type substrate have been shown in Figure 1.9. This was previously used to discuss primarily the quantum size effects in the silicon substrate that cause an effective increase in the equivalent dielectric thickness for the capacitance and for the channel inversion charge. The typical effect of a polysilicon gate as compared with a metal gate can be seen by comparing curves (c) and (d) with curve (b) which includes only quantum confinement effects. The capacitance values represent charge per unit voltage, so the differences between the curves (c) and (d), and curve (b) represent reductions in transconductance or channel charge that would occur due to polysilicon depletion. If the polysilicon doping density could be increased without limit, the polysilicon depletion effect could be minimized. However, from literature reports, it appears that it will be very difficult to achieve electronically active doping densities much above the $1\times10^{20}/\text{cm}^3$ for *n*-type silicon and above the mid to upper $1\times10^{19}/\text{cm}^3$ for *p*-type polysilicon. As can be seen from Figure 1.9, this represents a significant degradation in current drive capability for devices beyond the 90-nm node.

To overcome the polysilicon depletion effect, there are two approaches: (1) attempt to very heavily dope the polysilicon with values near $1\times10^{21}/\text{cm}^3$ needed, or (2) replace polysilicon with a metal gate. The most fruitful approach appears to be the metal gate approach, which is at this time being extensively pursued by the industry. For metal gates, two different metals with two work functions (WF) are needed—one for *n*-channel and one for *p*-channel devices. This is required to provide appropriate metal WFs to essentially replace the WFs of n^+ and p^+ polysilicon. Metals are needed with WFs near the conduction and valence bands of the underlying silicon.

For a direct replacement for n^+ and p^+ polysilicon, WF values of 4.1 and 5.2 eV are needed. For possible metal gates, a wide variety of WF values exist in the elemental metals ranging from below 3.0 eV for metals, such as Li, Rb, Sr, Cs, Ce, Pr, and Eu; to over 5.0 eV for metals such as Ni, Se, Rh, Pd, Ir, Pt, Ru, and Au. However, most of the elemental metals have significant problems when used as gate contact materials over either SiO_2 or the most promising high-*k* dielectrics such as the Hf silicates. The low WF metals tend to be too reactive, while the high WF metals tend to have adhesion problems. The latter adhesion problem can be controlled by the use of capping layers and bilayers of metals such as RuTa, RuMo, or TiPt have shown considerable promise for *p*-type MOS devices [21–23].

As a general class of materials, the metal nitrides and metal silicides are much more stable than the elemental metals and considerable research has been done on such compounds for possible metal gates. Some of the most promising of these are TaN [24], TiN [25], TiAlN [26], TaSiN [26], and HfN [27]. Metal silicides have been extensively used as source/drain contact materials for CMOS, so the possible extension of this technology to MOS gates is an attractive approach. For such an application, the metal silicide must extend completely to the gate dielectric and such a silicide process has been referred to a "full silicidation" (FUSI) process [28].

The threshold voltage of a MOS device depends not only on the fundamental or intrinsic WF of the metal gate, but also on the presence of any possible charge dipole layers within the silicon/dielectric/metal gate structure. Such dipole layers are possible at either of the dielectric interfaces or at an internal interface between a high-*k* dielectric and an interface oxide or nitride layer. Such dipole charge layers are known to pin the Fermi level for most metal-semiconductor interfaces near the midgap energy. For an MOS gate contact, such dipole layers will result in some effective WF that differs from the intrinsic gate WF. Evidence for such dipole layers have also been found for many metal-high-*k* interfaces. Typically, such dipole layers result in an effective WF more toward the midpoint of the Si bandgap and away from the band edges.

Some of the most promising metal gate materials are summarized below:

A. *p*-MOS (desired WF ~5.2 eV):
   Pt (5.4 eV), Ru (5.2 eV), Ti (4.6 eV), Mo (4.7 eV), TaAlN (4.9 eV), TiN (4.8 eV), CVD TaN (5.1 eV), and B doped FUSI NiSi (5.1 eV)
B. *n*-MOS (desired WF ~4.1 eV):
   Ta (4.2 eV), Ti (4.2 eV), TaSiN (4.2 eV), and As doped FUSI NiSi (4.6 eV)

As of this review, the most appropriate metal gates and integration approach has not been determined. However, the future of MOS devices most likely involves a return to the dominant use of metal gates.

1.4.2 Channel Doping Issues

The proper selection of substrate doping and engineering of the doped layers beneath the gate are key to optimum performance of the MOSFET device with respect to important parameters, such as threshold voltage, peak channel mobility, DIBL, subthreshold current slope, and drain-to-source punch-through voltage. Many of the issues regarding channel doping can be understood with reference to Figure 1.18, which illustrates the important charge regions under the channel. First, and perhaps most important, there are depletion regions under the gate (represented as $X_{dm}(y)$) and depletion regions surrounding the source/drain junctions (represented as X_{sd} and X_{dd}). The doping density under the channel controls the depth of all these depletion regions with larger doping densities required to obtain thinner depletion regions as device dimensions are scaled. Of primary importance is the requirement that the junction depletion regions do not overlap under the gate or else junction, punch-through will occur and large drain-to-source currents can flow. As the junction built-in voltage does not scale with device scaling, this requirement becomes more difficult to satisfy with scaling.

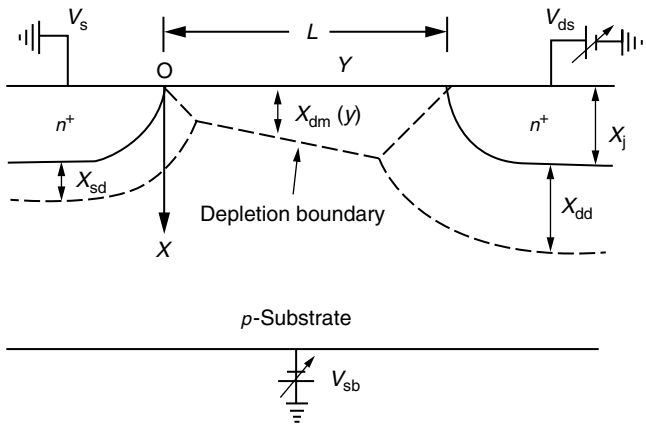


FIGURE 1.18 Charge sharing model for V_T reduction and other short-channel effects. (From Arora, N., *MOSFET Models for VLSI Circuit Simulation*, Springer, New York, 1993.)

Controlling short-channel effects are very critical to the ability to properly scale MOS devices. The most important of these are: (1) threshold voltage reduction, (2) DIBL effect, and (3) subthreshold current slope parameter. While these parameters are frequently expressed in terms of complex function of device parameters, such as junction depths, doping, oxide thickness, etc., they are primarily functions of the two-dimensional geometry shown in Figure 1.18. The dotted lines within the depletion regions are intended to represent the regions of charge controlled by the various electrodes (gate, source, and drain). One way to look at the drain voltage dependence of V_T is that part of the underlying depletion region is controlled from the source and drain, and not from the gate electrode. The fraction of charge controlled by the gate depends on the width of the depletions and junction depth relative to the gate length L. If the same relative size of the junction depths and depletion regions is maintained at each technology generation, then the relative importance of short-channel effects should remain the same. To achieve this, the depletion region widths must be scaled in the same manner as other device dimensions. Another way of viewing the drain voltage dependence of V_T is that the potential under the center of the gate is controlled not only by the gate charge, but is also influenced by the source and the drain depletion region charges. Adding more charge at the depletion regions will shift the potential under the gate, resulting in more channel charge for the same gate voltage.

Drain induced barrier lowering and changes in V_T arise from similar physical effects. At zero drain bias, there will be some lowering of V_T due to the source/drain charges. As the drain bias is increased, more charge at the drain depletion region will result in a larger effect and a further reduction in threshold voltage; hence a voltage-dependent threshold voltage or DIBL. This again is primarily a geometrically determined parameter depending on the junction depths, oxide thickness, and depletion layer depth relative to the channel length. Low DIBL requires that the field lines from the drain terminate on charges other than those in the channel. This is determined by the degree to which the gate is electrically shielded from the drain. Maintaining a fixed relative geometry is again the key to control the DIBL at small device dimensions. Each 10 mV of V_T reduction caused by DIBL will increase the I_{off} current by two to three times. However, DIBL does increase I_{dast} as well as increasing transconductance of the device. An optimum tradeoff occurs for a DIBL that is approximately 10% of V_{DD}.

The channel doping densities and doping profile must be selected to obtain a proper device threshold voltage and to control short-channel effects. Some of the implications of this, with respect to the doping of the scaled devices, is now explored and discussed. To control depletion layer punch-through, the channel must be sufficiently heavily doped under the gate region. There are several possible ways this can be achieved. First, the channel could just be uniformly doped with some large doping density in the region identified as the substrate body in Figure 1.7. However, since the channel mobility is degraded by a large bulk impurity density, there are reasons for desiring a lightly doped region just under the gate where the inversion channel exists. This can be accomplished by use of a retrograde channel doping such as shown in Figure 1.19, where a lightly doped region is shown directly under the gate with a more heavily doped region deeper into the substrate. This is frequently referred to as a super steep retrograde doping profile. Such a profile also gives added flexibility with respect to threshold voltage control, since the threshold voltage can be controlled somewhat by adjusting the depth of the lightly doped layer.

The super steep retrograde doping profile, however, does not provide sufficient flexibility to achieve all the desired goals of the channel doping for short-channel devices. For example, the doping density needed to control punch-through, and limit the size of the junction depletion layers can be too large to give a desired low threshold voltage even when the heavily doped layer is located some distance under the gate oxide. In such cases, it has been found highly desirable to employ another doping technique that allows a non-uniform lateral doping density under the gate. This is the pocket implant technique illustrated in Figure 1.20. This is also frequently referred to as a superhalo doping. By the use of shallow angled implants, using the gate as a mask, heavily doped pockets can be formed in the substrate adjacent to the source/drain contact implants. These heavily doped pockets can provide the heavily doped regions needed to terminate the junction depletion regions and control punch-through. However, they can leave a more lightly doped substrate under the center of the gate region, so that an acceptable threshold voltage can be achieved. For greatest flexibility, this should be combined with the super steep retrograde doping

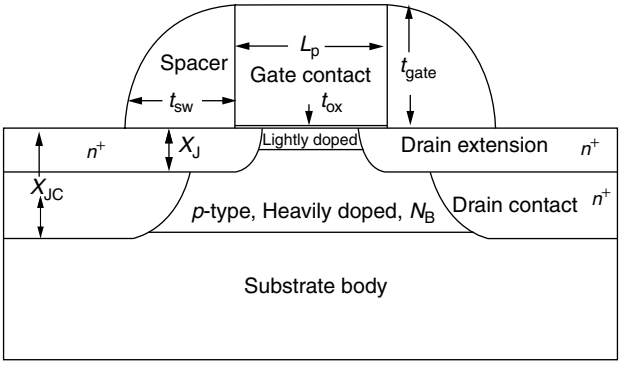


FIGURE 1.19 Channel structure and doping for super steep retrograde channel doping.

profile. In this manner, one can somewhat separate the problem of punch-through control from that of threshold voltage control. However, this becomes more and more difficult with dimensional scaling because of the large doping densities required, and because of the small dimensions over which the doping profiles need to vary by large factors.

To a first order approximation, the doping density for a given depletion width of the drain-to-substrate p–n junction with the drain side, which is very heavily doped is given by

$$N_B = \frac{2\varepsilon_s(V_j + V_{bi})}{qW^2}F \tag{1.32}$$

where N_B is the light side doping density, V_j is the applied junction bias voltage (maximum of V_{DD}), V_{bi} is the built-in junction potential ($\sim$0.9–1.1 V for $N_B > 10^{17}/cm^3$), W is the width of the depletion region, and the factor F accounts for the curvature of the p–n junction. For a planar junction, $F=1$; for a cylindrical junction where R_j is the radius of the p–n junction ($R_j \cong 0.65X_J$), F is given by

$$F = \frac{2[W/(W+R_j)]^2}{\ln\left(\frac{W+R_j}{R_j}\right) + \left(\frac{R_j}{W+R_j}\right)^2 - 1} \tag{1.33}$$

For shallow junctions, the cylindrical junction approximation is more accurate than the planar junction approximation. To prevent the depletion layer punch-through, W must scale with feature size

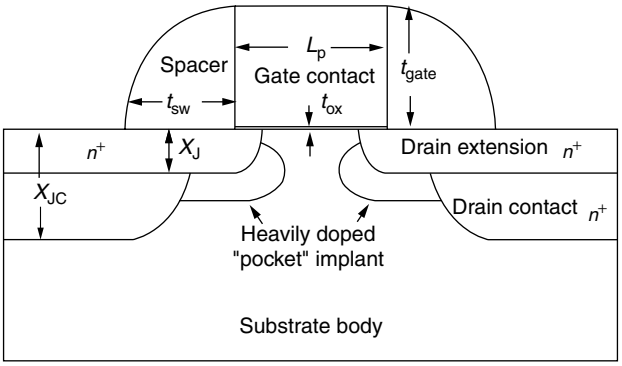


FIGURE 1.20 Example of heavily doped pocket implants for punch-through control.

and scaling W means that doping density N_B must scale somewhat faster than the inverse feature size, since $V_j + V_{bi}$ does not scale as fast as feature size.

Figure 1.21 shows the required doping densities according to these equations as a function of technology generation from the 90- to 16-nm node. Using the ITRS projections for dimensions, the factor F is approximately constant for each technology generation, depending only on the ratio of W to R_j Two potential limits are shown for the depletion layer: one for $L_{eff}/4$ and one for $L_{eff}/2$. The first giving a highly desirable limit, while the latter providing probably an upper limit on the acceptable value of the depletion layer width. For the two cases, $F = 0.806$ and 0.885, respectively, so the junction curvature does not give very much smaller limits than a planar junction with the values being only 12%–20% lower in value as can be seen in Figure 1.21. Also shown in the figure is a "mean" curve representing the geometrical mean of the $L_{eff}/4$ and the $L_{eff}/2$ curves for the case of a cylindrical junction. This can be taken as typical of the desired peak doping densities to control the junction depletion region depth with future scaling. While the $L_{eff}/2$ value might seem too large, the bulk of the depletion region will be confined between the two deep drain contact junctions of X_{JC} as shown in Figure 1.7. The spacing between these junctions can be two or more times the effective channel length. Thus, a depletion layer extension of $L_{eff}/2$ from these junctions is not unreasonable.

These required doping values are quite large and there are significant consequences of such large doping densities to be subsequently discussed. As a point of reference, Frank et al. have published a typical pocket implant doping profile for a 25-nm effective channel length device [29,30] and this is shown in Figure 1.22. Their peak doping density is somewhat above $1 \times 10^{19}/cm^3$ with a maximum depletion region depth of about $L_{eff}/2$. In comparing with Figure 1.21, this would correspond to the 90-nm technology node and this one data point compares very well with the $L_{eff}/2$ curves in Figure 1.21.

Large doping densities as shown in Figure 1.21 will give ever increasing drain-to-substrate leakage currents as the densities required for punch-through control are in the degenerate range and approach those of tunnel junctions. A model for junction tunnel current has previously been presented by Moll [31] as

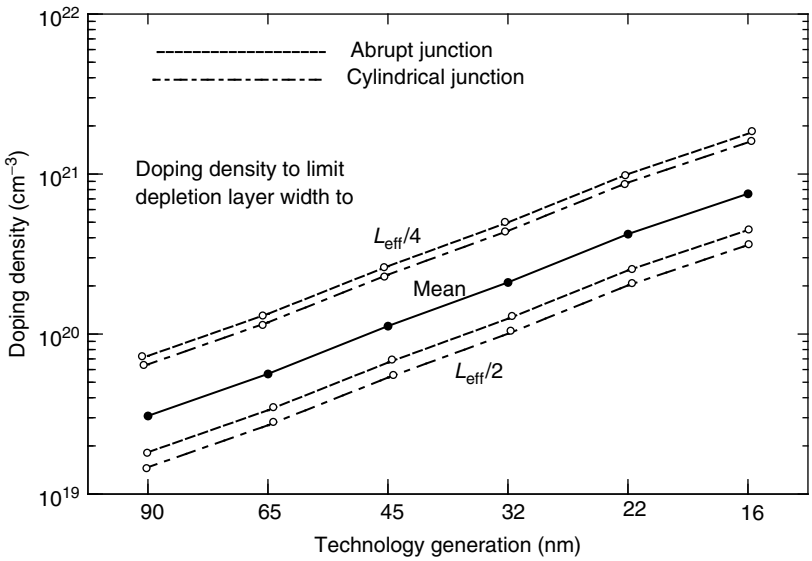


FIGURE 1.21 Doping densities required for punch-through control with different limits on the maximum width of the depletion layers.

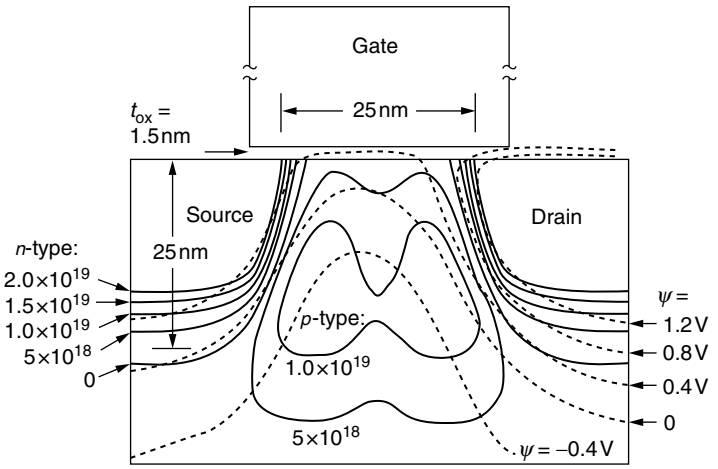


FIGURE 1.22 Doping concentrations for a 25-nm gate length device. (From Frank, D. J., R. H. Dennard, E. Nowak, P. M. Solomon Y. Taur, and H-S. P. Wong, *Proc IEEE*, 89, (2001): 259; Taur, Y. C. H. Wann, and D. J. Frank, *IEDM Tech. Digest.*, 789, 1998.)

$$J_t = \frac{\sqrt{2m^*}q^3EV_j}{4\pi^2\hbar^2\sqrt{E_g}}\exp\left[-\frac{4\sqrt{2m^*}E_g^{3/2}}{3q\hbar E}\right] \tag{1.34}$$

where m^* is the tunneling effective mass taken as $0.19m_0$, V_j is the applied junction voltage, E_g is the bandgap energy and E is the electric field taken here for an abrupt junction as

$$E \cong \sqrt{\frac{qN_B(V_{bi} + V_j)}{2\varepsilon_s}} \tag{1.35}$$

Calculations based upon this model are shown in Figure 1.23 for doping densities in the range of 1×10^{19} to $1\times 10^{21}/\text{cm}^3$ covering the range shown in Figure 1.21. Two curves are shown corresponding to the expected maximum device voltages. Shown across the top of the figure is the corresponding electric field over the range of 1–10 MV/cm. The values of current are similar to the values reported by Frank et al. [29], which includes some experimental data points. If one compares the doping densities in Figure 1.21 with the tunneling currents in Figure 1.23, it can be seen that there are large potential problems with junction leakage for the smallest device dimensions.

If one makes the same assumptions about allowable magnitudes of drain leakage current as for gate leakage current and assumes a somewhat optimistic scenario that the tunneling current can be confined to a spatial depth corresponding to the junction depth, then we can approximate

$$J_t \leq \frac{(2.5\times 10^{-4}\text{A/cm})}{X_J} = \frac{(2500\text{A/cm}^2)}{(X_J/\text{nm})} \tag{1.36}$$

This would indicate that the junction leakage current would exceed allowable limits around the 32-nm technology generation. Requirements for the control of depletion layer punch-through and tunneling currents probably limit the ability to scale bulk CMOS to dimensions somewhat less than the ultimate 2004 ITRS limits. Alternative device structures such as fully depleted (FD) SOI or double gate (DG) structures may be ultimately required. Such structures will be discussed in a section 1.5.2.

One approach for reducing the junction electric field that might come to mind is to dope the deep contacting junction more lightly, so that the junction is not basically a one-sided junction, but have the

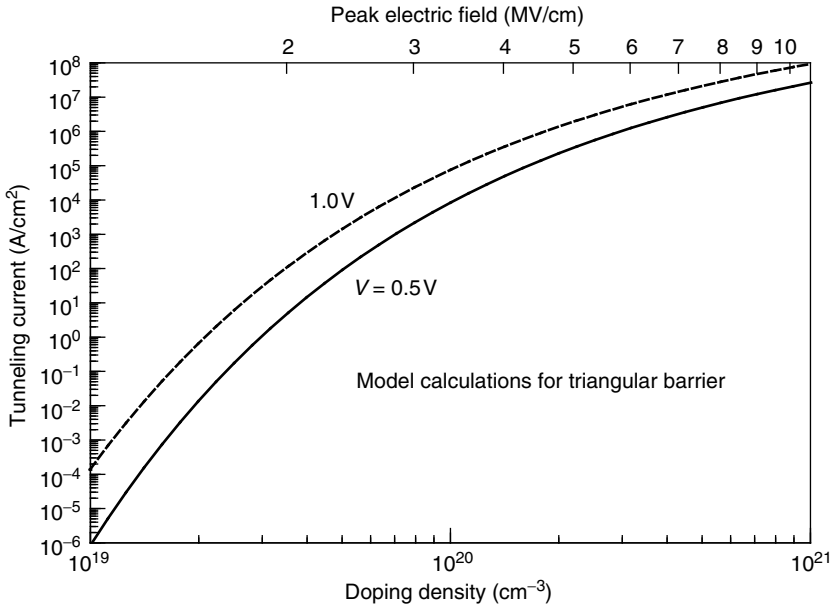


FIGURE 1.23 Estimated drain-to-substrate leakage current for various punch-through doping densities and peak electric fields.

depletion region extend more into the deep contacting junction. However, as will be subsequently shown, the deep contacting junction must be as heavily doped as possible in order to obtain an acceptable contact resistance to an external contact. Thus, there is little that can be done in practice to reduce the peak electric field by reducing the doping density on the source/drain contact side of the junction.

In addition to controlling the depletion regions from the source/drain junctions, the substrate doping is important in setting the device threshold voltage. In a given technology, one generally desires devices with various channel lengths from some minimum channel length to larger channel lengths. It is thus desirable to have a device threshold voltage for the minimum channel length that is not very different from that of a long-channel threshold device. For a long-channel device, if one assumes a retrograde doping profile as shown in Figure 1.19, then first order calculations can be made from the standard semiconductor equations regarding the relationship between threshold voltage, width of the lightly doped layer, and the doping in the heavily doped layer. To look at important trends and magnitudes, it will be assumed that the doping density in the lightly doped layer may be neglected and that the depth of the lightly doped layer is x_I, and that the depletion layer extends to a depth x_B into the heavily doped layer. The model for substrate charge is then

$$\rho = \begin{cases} 0 & \text{for } 0 < x < x_I \\ -qN_B & \text{for } x_I < x < x_I + x_B \end{cases} \tag{1.37}$$

Using this substrate charge density, one can evaluate the surface potential for the threshold-channel condition and evaluate the theoretical threshold voltage for a given bulk doping density and given depth of the lightly doped layer [32]. Inversely, one can determine the bulk doping density required for a desired threshold voltage, given the depth of the lightly doped layer and the other device parameters such as oxide thickness. One finally needs a gate WF, which can be conveniently taken as at the band edges of the semiconductor. One complication is the need to correct the standard equations for quantum confinement effects. This can be done with various published correction models or one can use the

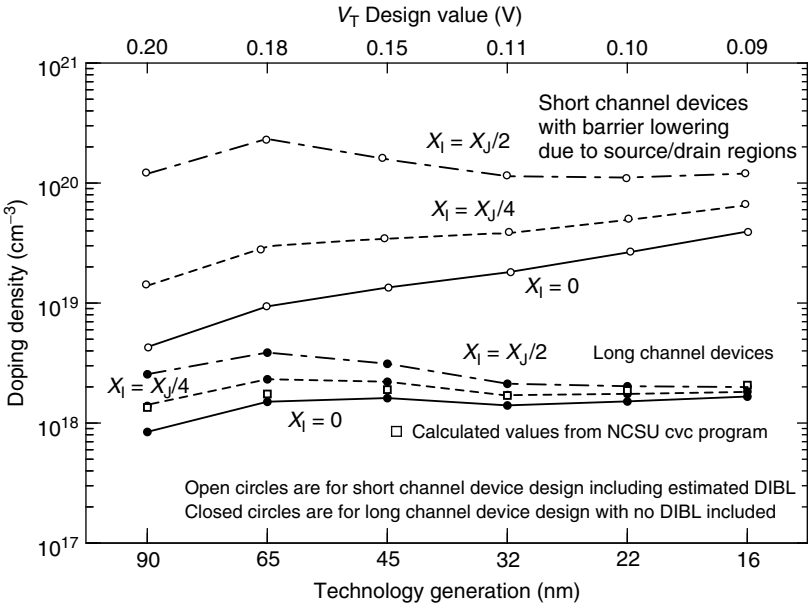


FIGURE 1.24 Doping densities for long- and short-channel devices needed with the retrograde doping model to achieve the design values of threshold voltage at each technology generation.

simplified approach of the ITRS to assume that these effects make the oxide thickness appear to be 0.4-nm thicker than the physical thickness.

Some of the important results of such an evaluation are shown in Figure 1.24 for the 90- to 16-nm technology generations using the dimensional parameters of the 2004 ITRS. Concentrate first on the three lower curves in the figure with solid circles as data points, which are the results for the calculation outlined above that would represent long-channel threshold voltages of the values seen on the top axis of the figure. The values are seen to be in the 1×10^{18} to $3\times10^{18}/cm^3$ range. The results are essentially independent of *n*- or *p*-channel type. Three curves are shown for lightly doped layer thicknesses of 0, $x_J/4$ and $x_J/2$ representing the most practical thicknesses of lightly doped surface layers. As seen by the curves, there is little difference in the required doping density for either the different technology generations or on the thickness of the lightly doped layer. For the calculations shown in the curves, quantum confinement effects were estimated by the technique of increasing the oxide thickness by 0.4 nm. Also shown in the figure are required doping densities for the device structures as determined by the NCSU cvc software, which includes a more detailed model for the quantum confinement effects, and which solves more exactly the device equations. These are the square data points and should be compared with the lower $x_I = 0$ curve. These points are just slightly higher in value that those obtained with the much simpler model and provide some additional justification of the simple model of increasing the effective oxide thickness by 0.4 nm to account for the QM effects. The NCSU cvc software cannot be used to compare the other curves, as it assumes a uniformly doped substrate. These values indicate why one cannot simply heavily dope the entire substrate to control the punch-through effect. The doping densities required for the proper threshold voltage are much lower than the densities shown in Figure 1.21 that are required for punch-through control.

As one decreases gate length from the long—to the short—channel case, moving toward a case as shown in Figure 1.22, several physical effects occur. First, the shielding of the source/drain depletion regions from the channel will become less effective and the threshold voltage will begin to decrease. Second, as the channel length becomes very small, the doping densities from any pocket implants will

begin to overlay and cause an increase in the doping density under the center of the gate. In the minimum channel length design, as shown in Figure 1.22, one can see considerable enhancement of the doping density under the center of the gate due to the overlap of the pocket implants. This increase in effective channel doping tends to increase the threshold voltage. Thus, with a pocket implant or halo structure one has two competing effects: one tending to reduce the threshold voltage and one tending to increase the threshold voltage. This can greatly delay the rolloff of threshold voltage with decreasing channel length and allows much shorter channel lengths than would be otherwise possible. As a function of decreasing channel length, it is not unusual for the threshold voltage to first increase and then decrease with decreasing channel length.

The decrease in threshold voltage with channel length has been extensively studied both theoretically and experimentally. Various models predict that the decrease in threshold voltage is an exponential function of distance of the source/drain from the center of the channel ($L_{\text{eff}}/2$) and depends on a length parameter that is a function of the geometry of the channel structure [33–38]. The other expected dependency is on the charge in the depletion layer, which gives the model form

$$\Delta V_{\text{T}} \propto \sqrt{V_{\text{D}} + V_{\text{bi}}} \exp(-L_{\text{eff}}/2\lambda) \tag{1.38}$$

If we assume contributions from both the source and drain then

$$\Delta V_{\text{T}} = V_{\text{DB}}(1 + \sqrt{1 + V_{\text{DD}}/V_{\text{bi}}}) \exp(-L_{\text{eff}}/2\lambda) \tag{1.39}$$

In this, it has been assumed that the source is at zero volts and the worst case drain condition is with the drain voltage at the supply voltage. The V_{bi} parameter is the equilibrium junction voltage (about 1 V), V_{DB} is the proportionality constant and λ is the exponential decay length. It should be noted that, not all authors use the same notation for the exponential term. In some cases, the λ term has just been called l [37] or Λ and in some cases the exponential term has been expressed as $-\pi L_{\text{eff}}/2\Lambda$ [29], which gives a different meaning to the length scale factor. The data of Wann et al. is consistent with the numerical values

$$\begin{aligned} V_{\text{DB}} &\cong V_{\text{bi}} (\cong 1 \text{ Volt}) \\ L_{\text{DB}} &\cong (t_{\text{ox}}(x_{\text{I}} + x_{\text{B}})x_{\text{J}})^{1/3} \end{aligned} \tag{1.40}$$

If one uses this model for the reduction in threshold voltage, this can be included in the simple model previously discussed for substrate doping, and one can evaluate the required increase in channel doping for obtaining a desired threshold voltage. The upper three curves in Figure 1.24 illustrate the results of including this model of channel threshold voltage reduction combined with the requirement at minimum channel length to have the same threshold voltage as for the long channel case. In other words, if the channel doping densities are as given by the upper three curves under the center of the channel for a minimum channel length device, the channel reduction effect will be offset by the increased voltage due to the increased doping. The required doping densities of the upper three curves are in the range that one might expect to achieve with the overlapping doping densities of a pocket implant.

A very accurate design of a pocket implant and the substrate doping density profile requires two-dimensional computer simulations. However, the above first order calculations illustrate that with a properly designed pocket implant, with an increased doping density under the channel for a minimum channel length device it may be possible to offset most of the DIBL effect and obtain the required low threshold voltages for scaled devices. If all dimensions are scaled together, the design problem is similar at each technology generation. However, the manufacturability of future scaled devices becomes more questionable, because the doping densities must increase and the dimensions over which the densities must change values become smaller.

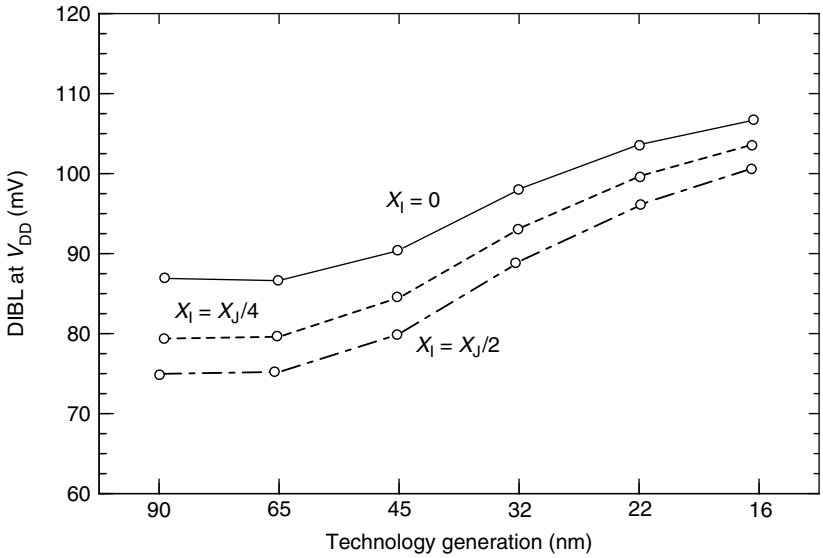


FIGURE 1.25 Estimated device drain-induced barrier lowering (DIBL) for nominal short-channel devices using a retrograde doping profile to set long-channel threshold voltage.

The model calculations also provide estimates of the value of DIBL that might be expected for such scaled devices. Assuming that DIBL is just the drain voltage manifestation of Equation 1.39, one can write

$$\text{DIBL} = V_{\text{DB}}(\sqrt{1 + V_{\text{DD}}/V_{\text{bi}}} - 1)\exp(-L_{\text{eff}}/2\lambda) \tag{1.41}$$

Plots of this are shown in Figure 1.25 for different assumptions on the width of the lightly doped surface layer. Controlling DIBL really means controlling the device relative geometry and if the relative geometry can be maintained as this modeling assumes, DIBL is expected to remain relatively constant and within acceptable limits (<100 mV in Figure 1.25).

A final important geometrical parameter is the subthreshold slope parameter. A model for this parameter has previously been given as Equation 1.4 and depends on the ratio of oxide thickness (corrected for QM effects) and the substrate depletion layer depth. For the retrograde doping model and the doping parameters in the previous figures, the predicted value of $\mathcal{S}$ is given in Figure 1.26. A small value of $\mathcal{S}$ (ideal value of 60) results from a deep depletion region and this is achieved for the long channel design. The increased substrate doping required at short channels degrades the slope parameter, but the resulting values are still very good. Again, if one can keep the relative geometry the same, the subthreshold parameter remains in good control, but the subthreshold slope is expected to degrade with short-channel length devices.

1.4.3 Source/Drain Contact Issues

The typical drain contact structure, shown in Figure 1.27, consists of the drain contact junction of depth X_{JC} and the drain extension of depth X_{J}. Although the initial purpose of the shallow drain extension was to reduce the peak electric field near the drain end of the conductive channel, and thereby reduce hot electron injection, as devices are scaled to lower voltages, the reduced field is not as important for this purpose. However, the shallow extension is required to reduce short-channel effects, as discussed in the previous section, while the deeper contacting junction is necessary to accommodate the thickness of the silicide contact layer and to minimize junction leakage. X_{JC} should not be larger than required to meet these conditions, since a larger value requires a thicker oxide spacer and a longer tab extension with

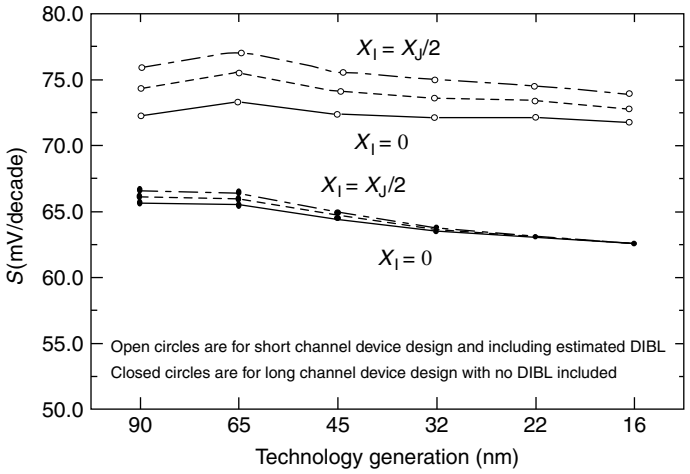


FIGURE 1.26 Estimated subthreshold slope factors for long and short-channel devices.

increased resistance in order to minimize short-channel effects. The source/drain resistance design should minimize both parasitic effects and contact resistance [39–42].

There are three major components of source/drain resistance: (1) an accumulation layer resistance due to the gate overlap with the drain region, (2) a spreading resistance as carriers spread from the channel into the X_J junction, and (3) a bulk junction resistance due to the length of the drain extension over the X_{JC} and silicide contact. The latter two components are especially significant with regards to scaling to smaller device dimensions. Referring to Figure 1.27, R_{tab} the resistance of the drain extension can be estimated as

$$R_{tab} \cong \frac{\rho \ell_{tab}}{W X_J} \quad \text{or} \quad R_{tab} W \cong \rho_{sc} \ell_{tab} \tag{1.42}$$

where ρ is the resistivity of the drain extension, ℓ_{lab} is the length of the drain extension, and ρ_{sc} is the sheet resistance of the drain extension. Without employing a two-dimensional analysis, a reasonable estimate for the tab extension length is $0.5L_P$ or half the physical gate length. For a constant sheet resistance, the $R_{tab}W$ product is then expected to scale as the device dimensions.

Several authors have studied the spreading resistance problem [39,40,43] and the analytical approximations from these studies can be written as

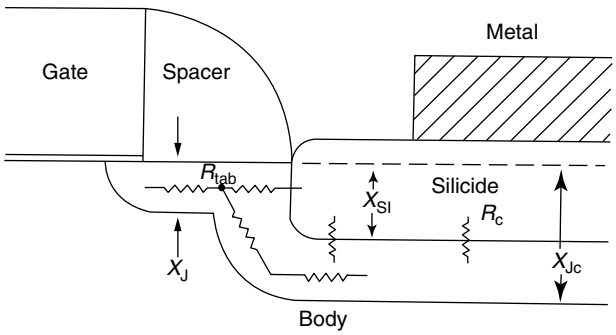


FIGURE 1.27 Illustration of the major components of source or drain resistance.

$$R_{sp} \simeq \frac{2\rho}{\pi W} \ln(\beta X_J / X_C) \tag{1.43}$$

where X_C is the inversion accumulation layer thickness and β is a constant that varies from 0.37 to 0.90 depending on the method of derivation. For the purpose of this discussion, an intermediate value of 0.58 [43] will be used. This resistance is to be added to the bulk extension resistance of Equation 1.42 to give the total extension resistance of:

$$R_{tab} W \simeq \rho(\ell_{tab}/X_J)\left[1 + \frac{2X_J}{\pi \ell_{tab}} \ln(\beta X_J / X_C)\right] = \rho(\ell_{tab}/X_J) F_{tab} \tag{1.44}$$

The spreading resistance increases the extension resistance by the quantity in brackets in the above equation (factor F_{tab}).

In order to estimate the magnitude of this term, a typical channel inversion/accumulation layer thickness must be estimated and this depends on the surface field or the surface inversion/accumulation layer density. For a typical oxide field of approximately 1×10^7 V/cm, an inversion/accumulation layer density of around $2 \times 10^{13}/\text{cm}^2$ is expected. The work of Stern can then be used to estimate an average inversion/accumulation region thickness of about 2.5 nm [5,6]. The quantity in brackets F_{tab}, in the above equation can then be expected to range from about a factor of 2.08 at the 90-nm node to about 1.0 at the 16-nm node.

Any component of the source/drain resistance must be compared with the channel resistance, R_{sat}, which in the on-state can be defined as:

$$R_{sat} = V_{DD}/I_{dsat} \tag{1.45}$$

Circuit simulations show that when the external source and drain resistances are equal to about 10% of this value, the saturated drain current will be reduced by about 8%. Thus, an acceptable value of resistance will be estimated as one, where the source or drain component is less than 1/20 of this value (5% contribution for each source and drain component). Using Equation 1.9 for the saturated drain current gives

$$R_{sat} W = \frac{V_{DD}}{v_{sat} C_{ox} (V_{DD} - V_T) F_I} \cong \frac{t_{civ}}{0.85 v_{sat} \varepsilon_{ox} F_I} \tag{1.46}$$

In the second form of the equation, the approximation $V_T/V_{DD} = 0.15$ has been used. If the factor F_I were an exact constant, then the $R_{sat}W$ product would be expected to scale directly with the capacitance EOT. For the device model previously discussed, it is found that F_I varies from about 0.65 to 0.80 for *n*-channel devices over the 90- to 16-nm nodes and varies from about 0.46 to 0.67 for corresponding *p*-channel devices.

Calculated results for $R_{sat}W$ are shown by the upper four curves in Figure 1.28 for the 90- to 16-nm technology nodes. Two curves are shown for both *n*- and *p*-channel devices. The open circle data points are calculated from the device model of Equation 1.9 for I_{dsat}, while the solid data points come from using Equation 1.45 and the 2004 ITRS values of I_{dsat} The two approaches give similar trends and magnitudes, with the 2004 ITRS current values giving slightly larger values. Also the values are higher for *p*-channel devices than for *n*-channel devices because of the lower saturated drain currents for *p*-channel devices. The general trend is for $R_{sat}W$ to become slightly lower for smaller device dimensions.

These values and Equation 1.44 can now be used to estimate required values of the extension junction resistivity. Using the 10% of channel resistance requirement leads to

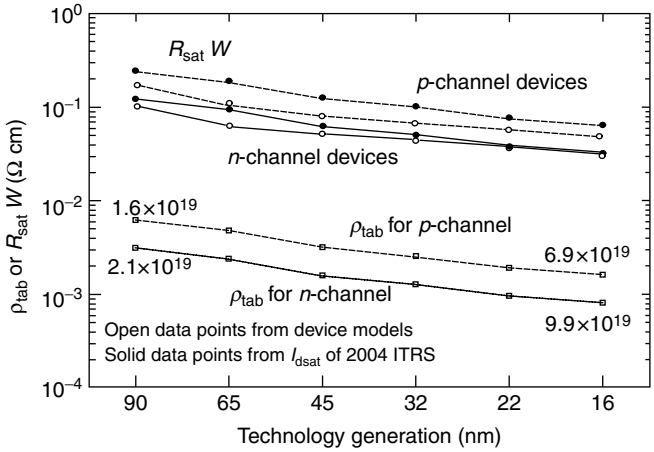


FIGURE 1.28 Estimated $R_{sat}W$ values and resulting contact tab sheet resistances for scaled metal oxide semiconductor devices. Estimated for approximately a 10% reduction in saturated drain current.

$$\begin{aligned}\rho(\ell_{tab}/X_J)F_{tab} &< \frac{1}{20}R_{sat}W \\ \rho &< 0.025R_{sat}W\end{aligned} \tag{1.47}$$

In the final expression l_{lab}/X_J is assumed to be constant (approximately equal to 1.0) and the worst case value of $F_{tab}(2.0)$ has been taken. This limit on extension junction resistivity is shown as the lower two curves in Figure 1.28 for n- and p-channel devices. Also shown in the figure are values of doping density corresponding to these resistivity values at the end points of the technology nodes. For example, for a p-channel MOSFET, the estimate is that the drain extension must be doped to at least $1.6\times10^{19}/cm^3$ for the 90-nm node and at least $6.9\times10^{19}/cm^3$ for the 16-nm node. Corresponding values for n-channel devices are 2.1×10^{19} and $9.9\times10^{19}/cm^3$. If drain extensions are formed by ion implantation with approximately Gaussian profiles, then the peak doping densities would be some 2–3 times the average values. In terms of sheet resistances, the corresponding limits are $1500\Omega/cm^2$ to $2400\Omega/cm^2$ for n-channel devices and $3000\Omega/cm^2$ to $4800\Omega/cm^2$ for p-channel devices. The required doping densities for the drain extensions are not too severe. However, the technology for producing junctions as shallow as needed for scaled devices is a critical technology that must be developed.

A final important resistance contribution comes from the deep contacting junction of depth X_{JC} and the contact resistance between the silicon and silicide layer. The most important component of this resistance is the contact resistance between any top metal or silicide layer to the silicon. A model for this resistance is [41]

$$R_C = \frac{\rho_{sc}}{WL_c'} \tag{1.48}$$

where ρ_{sc} is the contact resistance (in Ω/cm^2) and L_c' is an effective window length for uniform current flow with negligible resistance contribution from the sheet resistance of the X_{JC} junction. The effective window length can be expressed as:

$$\begin{aligned}L_c' &= L_c\tanh(L_c/L_t) \\ L_t &= \sqrt{\rho_{sc}/R_{sd}}\end{aligned} \tag{1.49}$$

where L_c is the physical contact length, L_t is known as the transfer length and R_{sd} is the sheet resistance of the underlying silicon comprising the X_{JC} junction.

The effective window length has two limiting cases of $L_c' \simeq L_c$ for $L_t \gg L_c$ and $L'_c \simeq L_t$ for $L_t \ll L_c$. In all cases $L_c' \leq L_c$ and therefore,

$$R_c \geq \frac{\rho_{sc}}{WL_c} \tag{1.50}$$

A good device contact design will have $L_c > L_t$. However, there is little advantage in making L_c larger than about $2L_t$. If $L_c = L_t$, the contact resistance is 30% larger than the right hand side value and for $L_c = 2L_t$ the contact resistance is only about 8% larger than the right hand side. Thus, making L_c larger than about $2L_t$, does not appreciably lower the contact resistance, but adds additional junction capacitance.

It will be assumed here that these simpler approximations apply and that the right hand side of Equation 1.50 provides a good estimate of the contact resistance. This requires that the contact junction have a sheet resistance satisfying the equation:

$$R_{sd} < \rho_{sc}/4L_c^2 \tag{1.51}$$

With these reasonable simplifying assumptions, in order for the contact resistance to contribute less than 5% of the device resistance one must have

$$R_c W \simeq \frac{\rho_{sc}}{L_c} \leq \frac{1}{20} R_{dsat} W \Rightarrow \rho_{sc} \leq 0.05(R_{dsat} W)L_c \tag{1.52}$$

For the estimates of device resistance in Figure 1.28, and estimated values of contact length, limits on the allowable contact resistance can be obtained. For this, we can estimate the contact length as approximately two times the MPU half-pitch, which is the same as the technology node.

Figure 1.29 shows the resulting required contact resistance values from this estimate for technology generations from 90- to 16-nm using the 2004 ITRS estimates of parameters and of saturation current. The values are essentially the same as those presented in the 2004 ITRS. The values are seen to be somewhat lower for *n*-channel devices as expected due to the larger saturated current of an *n*-channel device. The values range from the around 1×10^{-7} to the upper 1×10^{-9} Ω-cm^2. These are very low

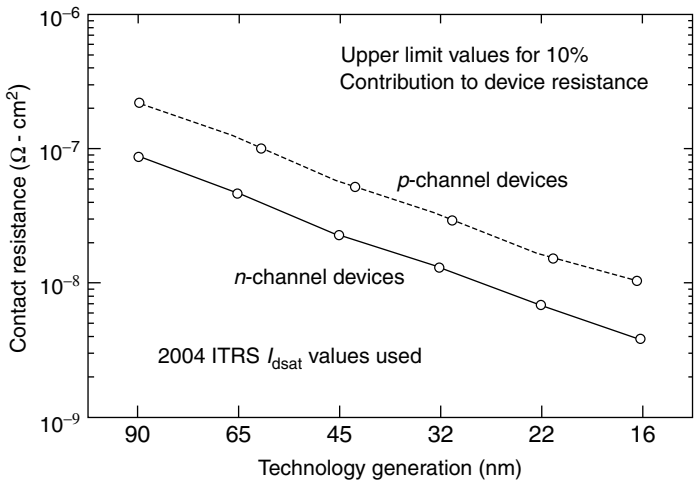


FIGURE 1.29 Estimated allowable source/drain contact resistances for scaled devices.

values and achieving such low contact resistances remains one of the major challenges of future MOS devices contacting structures.

To carry the analysis one step further, the contact resistance is typically assumed to be due to tunneling through the barrier between a metal (or silicide acting as a metal) and a heavily doped semiconductor. For such a model,

$$\rho_{sc} \propto \exp\left[\frac{4\pi\sqrt{\varepsilon_s m^*}}{h}\,\frac{\phi_B}{\sqrt{N_B}}\right] \tag{1.53}$$

In this, the important material parameters are the metal-semiconductor barrier height ϕ_B and the semiconductor doping density near the barrier N_B. A small contact resistance requires a small barrier height and/or a large doping density. For typical silicides on silicon, the barrier height is close to half the bandgap due to pinning of the Fermi level near the center of the bandgap. With a barrier height of around 0.5 V and for other silicon parameters, the predicted contact resistances for doping densities of around 2×10^{20} is around $1\times10^{-7}\ \Omega\text{-cm}^2$ [44]. This is consistent with the best contact resistances experimentally achieved with metal (or silicide) contacts on silicon. However, considerably lower values are needed for scaled MOS devices with projected values more than 10 times smaller required for the end-of-roadmap devices.

Considerable research work is underway on techniques for achieving the required low contact resistances. The most promising approaches involve the use of Si_xGe_{1-x} (SiGe) as an interface material between the silicon contact junction and the silicide or metal contact. The primary advantage of SiGe for this application is the lower bandgap, which can result in much lower contact resistances as predicted by Equation 1.53. Another advantage of SiGe is that the maximum doping density has been shown to be larger than that for pure Si and large doping densities can be achieved at low temperatures [45–48]. The optimum Ge concentration for low contact resistances appears to be in the range of 20%–30% and contact resistances as low as $1\times10^{-8}\ \Omega\text{-cm}^2$ has been demonstrated on both *n*- and *p*-type material [45–48].

Achieving the required values of drain extension and contact resistance is one of the major challenges of future MOS transistors. One modification to the basic device structure that helps these resistance values is the raised or elevated source/drain structure as shown in Figure 1.30. In this approach, the heavily doped drain contact is elevated above the original surface by using a selectively grown epitaxial layer. A silicide layer is formed on top of the epitaxial silicon layer either by selective deposition or conventional reacted metal. The structure has several advantages for aggressively scaled devices. First, the spacer oxide can be thinner than in conventional structures and this can greatly reduce the extension resistance as previously discussed. Second, the n+ contact layer can be sufficiently thick to contain the silicide layer without an excessively deep contacting junction or without excessive junction leakage. Third, the selective layer can potentially be heavily doped during growth to obtain a low contact

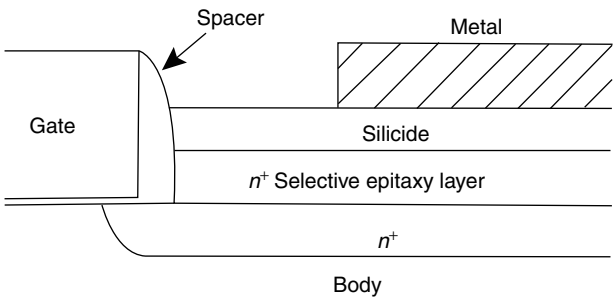


FIGURE 1.30 Illustration of the major components of an elevated source/drain contacting structure.

resistance. A final potential advantage relates to the possible mechanical stress that can be incorporated into a transistor from the SiGe contact structure. This is discussed in a subsequent section. However, a disadvantage of the elevated source/drain contact structure is that two selective epitaxy processes are required for both *n*- and *p*-channel devices since for CMOS both types of devices are required. The elevated S/D structure is expected to become essential for the ultimate scaling of MOS devices.

Also associated with the source/drain structure is a parasitic source and drain-to-substrate capacitance. To minimize this capacitance, the length of the contacts should be as small as possible. However, the overriding factor is probably the contact resistance so that one cannot reduce the contact length and the parasitic capacitance must be accepted. One means to reduce the parasitic capacitance is the use of SOI wafers and the advantages of such structures are explored in a subsequent section.

Because the source/drain resistances do not scale with the fundamental device dimensions in the same manner as device current, the source/drain resistances are major potential barriers to achieving the ultimate performance of scaled MOS devices. Creative structures and approaches will be required to extend scaling to the ultimate dimensions. The elevated source/ drain structure with heavily doped SiGe contact layers is one such approach.

1.4.4 Substrate and Isolation Issues

The substrate and isolation refers to those components of the integrated circuit that provide electrical isolation between the devices and prevents undesired device interactions such as latch-up in CMOS. Isolation has conventionally been achieved by the LOCOS structure. However, because of well-known scaling problems, such as large "'birds beak"' regions, new isolation techniques such as trench isolation have become essential for highly scaled CMOS. The 2004 ITRS assumes that shallow trench isolation (STI) will be the standard isolation technology for highly scaled CMOS. There do not appear to be major barriers to implementing implement trench isolation such as shown in Figure 1.31, although there are always technical challenges in performing fine line etches with high aspect ratios and uniformly filling with an insulator. The ability to implement improved isolation is essential to achieving the full benefits of increased scaling in terms of increased packing density. Latch-up control is expected to become less of a problem as operating voltages decrease and for voltages below 0.6 V should no longer be a problem.

1.4.5 Thermal Budget Issues

The scaling of MOS devices to ever thinner device layers necessitates going to lower thermal budget processing achieved by lower temperatures and/or shorter processing times. Some estimates of allowed thermal budgets can be made based upon expected device layer thicknesses. For example, the channel doping profile for a retrograde doping structures must not be deeper than the source/drain junction depth. Also the source/drain extension junction must be very shallow to control short-channel effects. These junctions must be considerably less than the feature size. If we make the assumptions that the peak

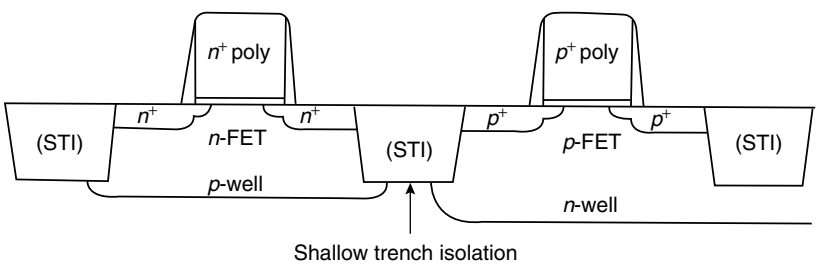


FIGURE 1.31 Shallow trench isolation.

doping densities are approximately 100 times the background doping density and that 50% of the junction depth can arise from diffusion, one can establish a maximum allowed value of the quantity $\sqrt{4Dt}$ which is approximately $0.23X_J$, where D is the impurity diffusion coefficient and t is the time at which diffusion occurs. This is essentially the amount of allowed diffusion, which ranges from approximately 4.7 to 0.8 nm at the 90- and 16-nm nodes.

D is typically modeled as a function of temperature with two parameters D_0 and E as:

$$D = D_0 \exp(-E/kT) \tag{1.54}$$

Figure 1.32 shows allowable times at different temperatures for the 90- to 16-nm technology nodes based upon the above defined amount of junction diffusion and previously given diffusion coefficient values [49]. Curves are shown for B and As impurities, but results for P are very close to B and results for Sb are very close to As. Results are shown for temperatures of 900°C and below as results for 1000°C are probably too short to be practical for the device dimensions in these generations. While there may be some variations from the values depending on the amount of allowed diffusion, these values do not include any transient enhanced diffusion rates, which may reduce the allowed diffusion times ever further.

These results illustrate the need for low thermal budget processes for ultra scaled MOS devices. For the same total process time, the temperature has to be reduced by more than 100°C in moving from the 90- to 16-nm nodes. Such low thermal budgets will also push processing toward single wafer, rapid thermal processing steps and will place much more emphasis on the control of transient enhanced diffusion processes. Limitations on the allowed processing temperature will become more of a concern as devices are scaled to ever smaller dimensions. This coupled with the very shallow layers with required very high doping densities will make ion implantation followed by annealing, more problematic for future devices. The times presented in Figure 1.32 do not appear to be sufficient to provide dopant activation and simultaneously to suppress transient enhanced diffusion. At some point other doping technologies will be required to achieve and maintain the required doping profiles. Possible alternative doping techniques include: (1) low temperature (800°C or less) epitaxy (or selective epitaxy), (2) diffusion from a doped glass source, (3) diffusion from a low temperature Ge/Si alloy which can be subsequently etched, (4) direct gas source diffusion in an rapid thermal processing (RTP) system, (5) gas immersion, laser diffusion (GILD), and (6) planar doping layers using atomic layer epitaxy. Of these many possibilities, low temperature

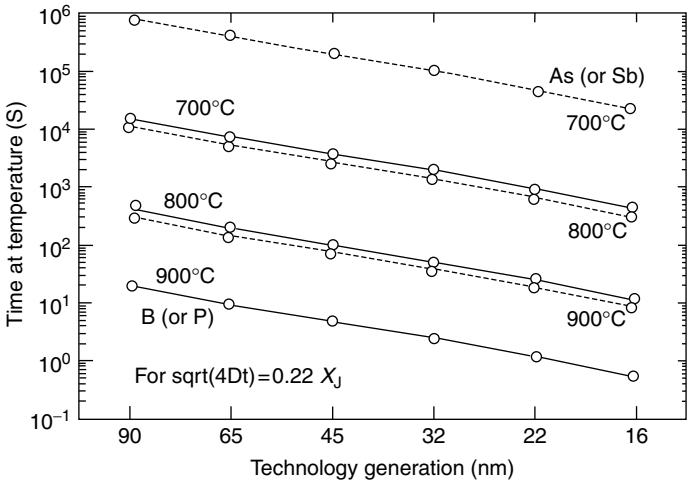


FIGURE 1.32 Limits to thermal budget for As and B with future device generations. Transient enhanced diffusion is not included.

epitaxy will probably find a major role as it will also probably be required for elevated source/drain structures for low resistance contacts as discussed in the previous section. With low temperature epitaxy, thin heavily doped layers can be achieved that are very difficult if not impossible to achieve by other techniques. However, with selective epitaxy, separate hard masking is required for making both *n*- and *p*-type S/D regions. GILD with projection masking (PGILD) does not require resists or hard masking and may become an important doping technology. Regardless of the doping technique, obtaining the required shallow junctions and very heavily doped regions will be a major challenge for scaled technologies. Silicon on insulator as discussed in the next section may also offer some advantages over bulk MOS in this regard.

1.5 Advanced MOS Device Concepts

While bulk CMOS has been the standard technology for many device generations, new substrate and device concepts may be essential to achieve the ultimate scaling of MOS devices. Some of the most promising approaches being explored are briefly discussed in this section.

1.5.1 SOI Substrates and Devices

The possible switch of substrate material to SOI at some future point may be essential for achieving the ultimate MOS device dimensions. Silicon on insulator devices have already been used in some applications for improved speed of operation. In SOI, the MOS devices are formed within a thin silicon layer formed by various techniques over an insulating layer (typically oxide or oxynitride). From the MOS device point of view, two types of situations can exist. If the semiconductor layer is sufficiently thick that the surface depletion layer under the gate terminates within the silicon layer, the device is a partially depleted (PD) device. If the silicon layer is sufficiently thin such that the silicon material under the gate is fully depleted (FD) before the inversion layer forms, the device is a FD device. In PD SOI, even though the material under the gate is only PD, the source/drain junctions can extend through the silicon layer giving reduced parasitic capacitances and higher operating speed for a given technology generation. Device isolation and latch-up control are simplified for SOI with fewer processing steps required. This can potentially offset the higher substrate costs and potentially lower yield. From a device viewpoint, there are some drawbacks to PD SOI due to a floating body effect. To eliminate such effects, a contact is required to the body increasing somewhat the area associated with a given transistor. However, a separate body contact can be of advantage in some logic approaches where threshold voltage switching is used with a separate body voltage.

A major advantage of an SOI substrate is that the depth of the contacting source/drain junctions can be controlled by the thickness of the silicon layer over the oxide layer. This applies to both FD and PD devices and is in one way, the shallow junctions required for deeply scaled MOSFETs can be achieved. For PD devices, this is the major advantage of an SOI substrate as the device characteristics and scaling properties are essentially the same as for the bulk MOS devices. Most of the potential advantages and limitations of FD SOI devices are very similar to DG MOS devices as discussed in Section 1.5.2, so a discussion of the FD MOS device is combined with the DG device. The discussion there will show that the feasibility of FD SOI devices is questionable at the 90 nm and below technology generations because of excessive short-channel effects.

One disadvantage of the SOI wafer is the power dissipation capability, which is less than a conventional Si wafer because of the poorer thermal properties of an oxide layer. For HP circuits with a high percentage of active devices, this can be a major disadvantage as power dissipation is becoming more of a limiting factor for scaling as discussed in section 1.3.2.

1.5.2 Multiple Gate MOS Devices

Considerable research and modeling has been performed in recent years on multiple gate MOS devices. One of the most extensively studied multiple gate devices is the DG device shown in simplified form in

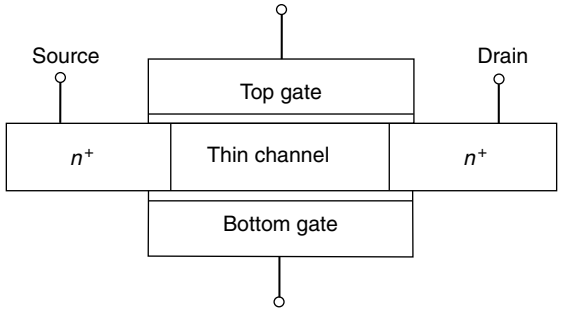


FIGURE 1.33 Simplified schematic of dual gate MOSFET structure. Wrap around gate device also has similar structure.

Figure 1.33. The structure consists of a thin conductive channel sandwiched between two control gates consisting of dielectric and gate contact material. Heavily doped source and drain regions extending slightly under the gates make contact to the inversion regions. The DG device is typically assumed to be sufficiently thin as to constitute a FD region within the semiconductor channel. A major advantage of the DG device is that the current drive can potentially be twice as large as that of a single gate device since an inversion layer can exist near both the gates. For very thin channels, these two charge layers can merge into a single conductive layer within the channel.

The practical fabrication of such DG devices is very difficult and no universally accepted technique has been developed for such structures. Some techniques attempt to fabricate such structures using a planar arrangement in which the channel length and width are in the plane of the silicon wafer. This has many advantages in that the W and L device ratio can be controlled by the lithography in much the same way as conventional MOS devices. However, forming the dual gates above and below a planar thin silicon channel is very difficult, especially in obtaining exact alignment of the two gates. Such techniques typically use some form of SOI layer for the thin silicon channel. Other fabrication approaches, move the plane of the conductive channel out of the wafer surface and into the vertical direction or perpendicular to the wafer surface. The so called "finfet" is one such example, where the width dimension of the channel is taken in the vertical direction [50,51] with the channel length parallel to the wafer surface. (Also similar to the DELTA FET [52]). The gate then typically wraps around the top of the channel. Variations of this approach are the "tri-gate" device, where the channel width is increased and the gate occurs on the top and two sides of the channel. Another approach is to place the channel in the vertical direction with current flow in the vertical direction from a bottom source contact to a top drain contact. In this approach, the gate can completely surround the channel giving the so called "gate-all-around" or "surrounding-gate" device. All of these multiple gate devices have somewhat similar properties and similar advantages and disadvantages. The dual gate structure will be discussed here as an example of all such multiple gate devices.

For the DG structure as in Figure 1.33 important properties are the control of threshold voltage and short-channel effects. This structure has been extensively studied theoretically [33–37,53] and the exponential decay factor for controlling short-channel effects has been shown to be

$$\lambda = \sqrt{\frac{\varepsilon_{si}}{2\varepsilon_{ox}}\left(1 + \frac{\varepsilon_{ox} t_{si}}{4\varepsilon_{si} t_{ox}}\right) t_{si} t_{ox}} \tag{1.55}$$

It has been reported that short-channel effects are acceptable, if $L_{eff}/2\lambda > 3$. If we use this value in Equation 1.39 to estimate threshold voltage reduction, this value is consistent with a threshold voltage reduction of about 0.12 V, where 0.1 V is sometimes taken as a limit on acceptable short-channel effects.

From the above limit, an equation for the upper limit on silicon thickness can be obtained as

$$t_{si} \leq \frac{2\varepsilon_{si} t_{ox}}{\varepsilon_{ox}} \left[\sqrt{1 + \frac{1}{18} \left(\frac{\varepsilon_{ox} L_{eff}}{\varepsilon_{si} t_{ox}} \right)^2} - 1 \right] \tag{1.56}$$

For a given effective channel length and oxide thickness, this can be used to estimate the allowed thickness of the silicon layer. However, before computing such values, some thought should be given to the oxide thickness term. This equation was derived without consideration of any second order effects such as Quantum size effects, which cause the inversion channel to not be located at the oxide interface, but some distance into the silicon. Also the derivation did not consider the possible use of high-k gate dielectrics. The latter effect can be taken into account by using the EOT value, since t_{ox} always appears divided by ε_{ox}. For oxide thickness, it is probably most appropriate to replace the t_{ox} term by the inversion capacitance oxide thickness value (t_{civ} as used here) to thereby correct for quantum size effects.

Figure 1.34 shows predicted limits on silicon thickness for the 90- to 16-nm technology nodes. Two limit curves are shown, one using t_{civ} and one using t_{ox} in Equation 1.56. While both show similar trends, the limit for t_{civ} is slightly smaller than for t_{ox} as would be expected. Regardless of which limit is used, the most striking results is the very small value of the required silicon thickness, which ranges from approximately 5–7 nm for the 90-nm node, to approximately 0.4–0.7 nm for the 16-nm node. However, these values are consistent with other published results. For example, it has been published in Ref. [29] that a 5-nm Si film with a 1.5-nm oxide can be scaled to a channel length of about 20 nm. These three values are all consistent with the 65-nm node data in Figure 1.34. Other researchers have suggested that the thickness of a Si layer may be somewhat thicker than shown in Figure 1.34 as determined by the above analysis. For example Doyle et al. suggest that an acceptable relationship for DG devices is $t_{si} < 2L_{eff}/3$ [54]. This limit is about a factor of 2 larger than the values seen in Figure 1.34. Such a value would extend the range of possible DG devices, but at the expense of larger short-channel effects.

Other curves shown in Figure 1.34 for reference are the effective channel length and the two possible values to use for oxide thickness. Also shown in the figure is a horizontal line at approximately 3 nm indicating that for Si thicknesses below this value quantum box effects must be considered. This arises because the width of the surface inversion layers is on the order of 3–4 nm in Si with holes having a slightly broader distribution than electrons [55]. When the thickness of a semiconductor layer

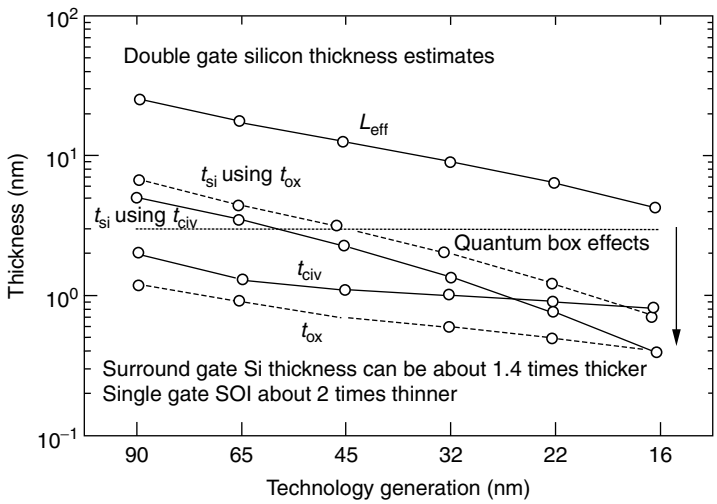


FIGURE 1.34 Thickness of Si for double gate (DG) MOSFET needed to control short-channel effects.

approaches this value, quantum size confinement effects will begin to exist. This will cause the energy level for the conduction electrons (or holes) to increase and for the required device threshold voltage to increase. Because of this, it has been projected that it will not be practical to use channel thicknesses very much below approximately 5 nm [29]. This is a very serious potential problem and makes the projections for using DG devices beyond about the 45-nm node very questionable.

In the previous section, the possibility of FD single gate MOS devices on SOI was discussed. In many ways, a single gate FD device on SOI looks like one half of a DG structure. If one splits the DG device along the center line parallel to the gate oxide and replaces half of the device by a thick oxide one has essentially a SG SOI device. It might then be expected that the FD SOI single gate device would have essentially the same limits as shown in Figure 1.34 for silicon thickness. However, the limits are not quite these, but are more restrictive. For a FD SOI device, there is an additional feedback path from the drain to the gate through the thick oxide at the bottom of the channel. This causes the short-channel effects to be more severe than for the DG structure. Several investigators have shown that the short-channel effects for single gate FD device occur at about twice the channel length as for the DG FD device. Thus, for single gate SOI devices the allowed silicon thickness can be taken as approximately one-half the values given in Figure 1.34. This makes the values very small and brings into question the feasibility of FD SOI devices at 90 nm and below. A major difference between bulk CMOS devices and SOI and the DG device is the ability to compensate for some of the short-channel effects by the use of overlapping halo implants with bulk CMOS. To achieve a similar effect with DG devices would require some type of lateral variation of channel impurity density in DG devices. This might be possible in vertical-channel devices where the vertical channel is etched from within a deposited semiconductor layer, but would seem to be very difficult to achieve in most DG devices structures.

The discussion has so far been on the DG device where much theoretical work has been done. For other multiple gate device structures, such as, the tri-gate, finfet, or surround gate device structure, the thickness of the allowed silicon structures may be slightly larger than for the DG structure. An estimate can be provided by the cylindrical FD device structure, where it has been shown that the natural scale length is about 30% smaller as compared with the DG structure. This means that the silicon thickness for a cylindrical gate device can be about a factor of 1.4 larger than for the DG structure.

The short-channel properties of MOS devices are controlled primarily by the $\alpha = L_{\text{eff}}/2\lambda$ parameter. This was discussed earlier for bulk MOS devices in connection with the threshold voltage reduction and DIBL. For FD devices, it has been shown that [35]

$$S \simeq (kT/q)\ln(10)\left[\frac{1}{1-2\exp(-L_{\text{eff}}/2\lambda)}\right] \tag{1.57}$$

This has also been shown to apply approximately to the surround gate device, which represents a limiting case of a multiple gate device. A plot of this equation is shown in Figure 1.35. For FD devices the long-channel value of subthreshold slope approaches the ideal value of approximately 60 mV/dec. As the figure shows, the value degrades rapidly for values of $L_{\text{eff}}/2\lambda$ less than about 3.0, which represents the limit used in estimated the allowed value of Si thickness in Figure 1.34.

The threshold voltage reduction due to short-channel effects can be estimated with the same equations as previously used (see Equation 1.39). This gives the predicted short-channel effect shown in Figure 1.36, which shows three curves for different drain voltages. Again the predicted threshold voltage decreases rapidly for $L_{\text{eff}}/2\lambda$ less than about 3.0. These theoretical curves are very similar to the published results obtained by much more complete two-dimensional numerical simulations of FD MOS devices. Although, there may be some uncertainty in the exact value of $L_{\text{eff}}/2\lambda$ at which a given value of threshold voltage reduction occurs, this curve should provide a reasonable first order estimate of the effect. Also shown as a vertical line is the limit used to establish an upper limit on Si thickness of FD devices and this is seen to occur at a threshold voltage reduction of about 0.1 V, which is probably a reasonable value when the actual threshold voltage must be on the same order of value.

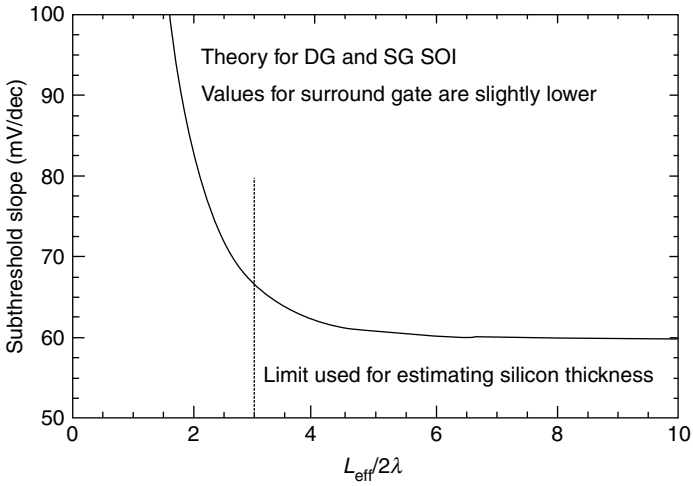


FIGURE 1.35 Variation of DG subthreshold slope factor on normalized effective channel length.

For FD devices, the long-channel threshold voltage has been estimated as [37]

$$V_T = \begin{cases} \dfrac{qN_A t_{si} t_{ox}}{\varepsilon_{si}} & \text{for single gate SOI} \\ \dfrac{qN_A t_{si} t_{ox}}{2\varepsilon_{si}} & \text{for double gate} \end{cases} \tag{1.58}$$

These represent simply the voltage required to fully deplete the Si layer. From this we can see that the accuracy of the threshold voltage depends on the accuracy with which the thickness of the Si layer can be controlled. For a 10% variation in threshold voltage, the Si layer thickness needs to be controlled to about 10%, i.e., for a 5-nm Si layer the thickness control must be about 0.5 nm [29]. Such requirements of tight thickness control coupled with the very thin layers will prove very difficult to meet for future devices.

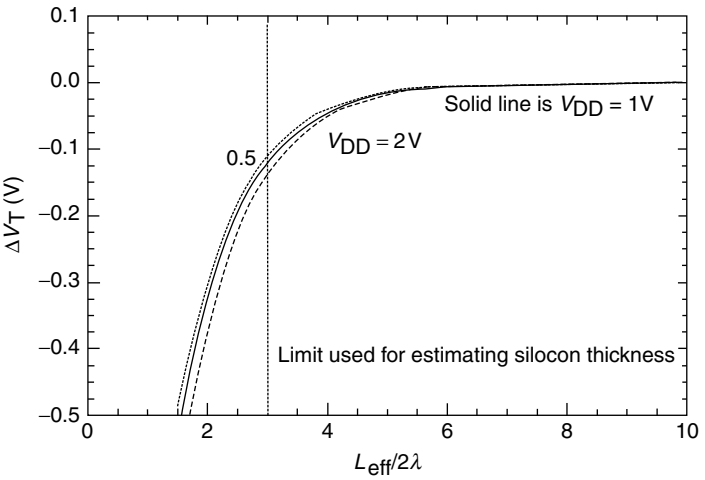


FIGURE 1.36 Variation of DG MOSFET threshold voltage on normalized effective channel length.

At present, the most promising techniques for producing SOI layers are SIMOX, Smart Cut and Bonded Wafers. All of these have great difficulty in obtaining such thin layers with the thickness control projected as needed for future MOS devices. Some fundamental breakthroughs are needed in fabrication thin layers for FD SOI devices as well as for multiple gate devices.

The most recently researched multiple gate devices, such as the finfet or tri-gate devices do not alleviate the requirement of ultra-thin Si layers for FD devices. For such devices, the thickness of the Si fin used in fabricating the devices must satisfy the same requirements as a DG device. This means that to control short-channel devices, the thickness of the Si layer must be considerable less than the channel length as indicated in Figure 1.34. In fact, the channel thickness must be four to five times less than the channel length. In most of the out-of-plane multiple gate device concepts, the channel thickness must be controlled by an etching process with this minimum width controlled by the lithographic process. This means that the channel length must be four to five times larger than the minimum lithographically defined dimension. This will greatly limit the current capability of such devices and may in fact completely offset any current enhancement advantage of such multiple gate devices. Another problem with out-of-plane MOS devices is that one does not have control of both L and W by the lithographic process, but one of these is controlled by the thickness of some Si layer. This makes it very difficult to adjust the W/L ratios as needed to compensate for the lower mobility of holes and to compensate for the delay times of various logic gates as well as adjusting this ratio for various analog applications. The W/L ratio can only be adjusted in steps by using multiple devices again offsetting some of the advantages of multiple gate devices. Other major problems with existing techniques for producing SOI material is the degradation in material quality with thin layers. Most experimental results have shown a degradation in carrier mobility when the SOI layer thickness is reduced below about 10 nm [29]. While it is uncertain if this is just a material quality problem or a fundamental problem with increased scattering processes, it remains a serious problem with existing SOI techniques.

For the reasons outlined above, multiple gate FD MOS devices face very formable obstacles in competing with bulk MOS devices. Controlling short-channel effects require very thin FD Si layers with respect to channel length and manufacturing techniques for producing such thin layers are not available with electrical properties comparable to bulk Si. Breakthroughs will be required in the manufacture of such multiple gate devices, if they are to become the mainstream MOS devices. A final summary of some of the multiple gate MOS concepts is shown in Table 1.3 along with some of the potential advantages and weaknesses.

TABLE 1.3 Summary of Some Multiple Gate FET Concepts

Concept	Tied Gates ($N_{gates} > 2$)	Side Wall Conduction (Finfet, Tri-gate)	Double Gate, Planar Conduction	Independent, Double Gate, Planar Conduction	Vertical Conduction, Wrap Around Gate
Advantages	Higher I_d and thicker fin	Higher I_d, improved S, and improved SC effects	Higher I_d, improved S, and improved SC effects	Improved SC effects	Improved SC effects and 3D integration
Particular strengths	Thicker Si body	Ease of integration	Bulk compatible and good Si thickness control	Electrically adjustable threshold voltage	Litho independent gate length
Potential weakness	Limited device width and corner effects	Fin thickness and shape	Limited width	Difficult integration and degraded S	Single gate length process integration

Source: Selected from 2004 International Technology Roadmap for Semiconductors, http://www.itrs.net/Links/2004Update/2004Update.htm

1.5.3 Transport Enhanced MOS Devices

One technique for enhancing the current capability of MOS devices at small dimensions is the application of appropriate mechanical strain to the silicon channel region [56–60]. Mechanical strain causes important changes in the internal band structure of the Si. Figure 1.37 illustrates some of the important changes in the bands [60]. For the conduction band, strain splits the sixfold degenerate energy bands along the 100 crystal directions into two groups of bands. With sufficient strain, all the electrons can be transferred essentially into the lower group of bands and if transport is along the appropriate crystal direction, the electrons will behave along the channel as if they had a smaller effective mass than for the unstrained Si. For the valence band, strain again splits the degenerate heavy and light hole bands into two bands with each band having a smaller effective mass than the heavy hold band in the unstrained Si. The most effective type of stress depends on the direction of stress relative to the MOS channel, and for *p*-channel devices should be compressive stress along the channel length dimension or tensile stress along the width dimension. For *n*-channel devices, the stress should by tensile stress along either channel direction.

The smaller effective mass of electrons or holes resulting from the splitting of the degenerate energy bands can enhance carrier transport along the channel. A very simple illustration is for the carrier mobility, which can be expressed in the most basic form as

$$\mu = \frac{q\langle\tau\rangle}{m^*} \tag{1.59}$$

where $\langle\tau\rangle$ is the mean time between scattering events and $m^\star$ is the effective mass. A smaller effective mass in this equation will result in a larger mobility. This is somewhat oversimplified as the mean time between scattering events may also depend on effective mass. However, for Si, the net result of appropriate mechanical stress is to give an enhanced low field carrier mobility and this has been experimentally

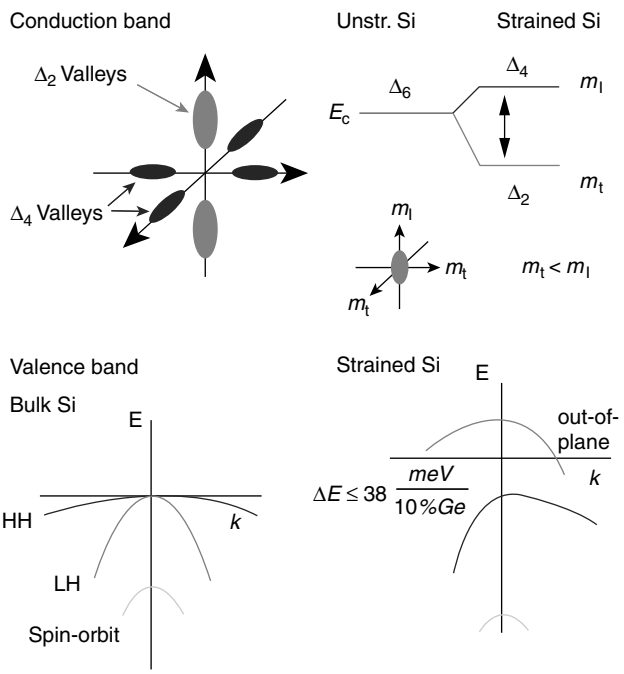


FIGURE 1.37 Effects of Si strain on energy bands. (From Ieong, M., B. Doris, J. Kenzierski, D. Rim, and M. Yang, *Science*, 306, (2004): 2057.)

verified for both electrons and holes. The mobility enhancement factor has been found to be as large as 1.83 for electrons and 1.53 for holes at low fields [56–60].

Just as important as mobility for determining current drive is the saturated drift velocity of electrons or holes. The saturated drift velocity arises physically because of the rapid generation of optical phonons when the carrier energy in a high electric field exceeds the optical phonon energy ($E_{op} \simeq 0.05$ eV for LO phonon). A somewhat oversimplified model for the saturated drift velocity predicts [61]

$$v_{sat} \simeq \sqrt{\frac{8E_{op}}{3\pi m^*}} \tag{1.60}$$

From this, one can see that a reduction in the effective mass should result in a larger saturated drift velocity. From the device model in Section 1.3.2, one can see that a larger low-field mobility and a larger saturated velocity will both contribute to a larger device current for all other parameters remaining fixed. The application of mechanical stress will likely become a standard technology to enhance device current as devices are pushed to ever smaller dimensions. Reported results indicate that the MOS drive current can be enhanced by 15%–25% with the use of mechanical stress.

There are several possible ways by which mechanical stress can be obtained in MOS devices. One approach uses a stress relaxed SiGe layer between a Si device layer and a Si substrate. When the SiGe layer is sufficiently thick, the stress will relax in the layer producing a SiGe layer with a lattice constant larger than that of pure Si. When a sufficiently thin Si layer is then grown expitxially on top of the SiGe layer, the resulting Si lattice will seek to match that of the SiGe layer giving a larger lattice spacing in the plane of the film and a Si film under tension. Other possibilities for stress include deposited films of various types that are under stress at room temperature, such as nitride films. For optimum *p*- and *n*-channel devices, the *p*-channel device should be under compression, while the *n*-channel device should be under tension. One attractive means of accomplishing this has been reported (by Intel) and consists of separate approaches for the two types of transistors [62,63]. For the *p*-channel device, a SiGe layer is selectively grown in the source/drain regions of the transistor. Since, the SiGe layer has a larger lattice constant, it tends to expand in the source/drain region putting the Si channel under uniaxial, compressive strain. For the *n*-channel devices a high tensile silicon nitride film is deposited over the top of the transistor resulting in a uniaxial, tensile strain in the *n*-channel device [62,63]. There are certainly questions as to how well such techniques will work for multiple-channel length devices, but strain optimization will probably be an important tool in the future device designer's toolbox.

1.5.4 MOSFETS with Other Semiconductors

Another possible approach to the enhanced transport is to move to another semiconductor such as SiGe alloys for the active region of the MOS devices. Other more exotic semiconductors such as InSb (with very high mobilities) may also be considered. These are certainly research issues to be pursued. However, one must not just look at carrier mobility, but at such parameters as saturated drift velocity which is probably more important than low field mobility. For Ge, the saturated drift velocities are known to be lower than for Si because of a lower optical phonon energy (see Equation 1.60). The use of SiGe with a small Ge concentration might give an enhancement from mobility before the loss due to lower velocity saturation occurs, but this is not obvious. For any move away from Si as the channel material, obtaining an acceptable gate dielectric will also be a formable task. Because of all the multitude of problems with any alternative semiconductor it is not clear that such concepts for enhanced transport will be practical in the real world or for cost competitive CMOS ICs.

1.5.5 Advanced Semiconductor Device Concepts

As Si CMOS devices are scaled to their ultimate limit in size, there is continued and growing interest is exploring new semiconductor device concepts that might have lower size limits or lower power

requirements that could continue the integration of electronic functions to even greater densities and higher performance. While there is no guarantee that such efforts will be successful, they must be pursued so that all possible avenues for extending the electronics revolution beyond the fundamental Si limits can be explored.

Viewed in the broad context, MOS devices are charge-based devices. A gate voltage controls the conductive charge in the MOS channel. As such, all charge-based devices are limited by the capacitance, voltage and current capability. This is captured succinctly in the CV/I term related to switching speed. All charge-based device concepts will have somewhat similar limitations and it is difficult to envision other materials and physical arrangements of charge-based devices competing with Si-based MOS devices both from a significant performance advantage and from a cost perspective. Thus, advanced device concepts for beyond Si MOS limits are probably best sought in device concepts that are not charge-based device concepts. Some of the most important avenues are

1. *Molecular and biological-based devices*. In these concepts the properties of individual molecules are used for transport and charge storage. Potential advantages are the small possible size of individual molecules, which might constitute a single device.
2. *Spintronics*. These concepts seek to use the properties of spin with its two quantized states to represent and process information. Potential advantages are the storage and manipulation of information on a single electron.
3. *Quantum interference devices*. Involves concepts for performing logic operations using the wave nature of electrons and the quantum interference effects between electron waves. Again the potential advantage is the potential size of devices.
4. *Phase change devices*. This broad classification involves using phase changes to store and process information such as, crystalline-amorphous, magnetic-nonmagnetic, or metal-insulator phase changes. Potential advantages are breaking the dependence on charge-based concepts.
5. *Optical switches*. Involves concepts to use photons for information processing and storage instead of electrons. Another non charge-based approach.

Much research is needed in any of these areas to bring forth alternative device concepts that can successfully compete with Si-based MOS devices. Some of the most promising non-MOS device concepts are summarized in Table 1.4. Some of these are still charge-based devices that have potentially smaller size

TABLE 1.4 Summary of Some New Device Concepts

Device Type	1D such as Carbon Nanotube	Resonant Tunneling Devices	Single Electron Devices	Molecular-Based Devices	Quantum Cellular Automata (QCA)	Spin-Based Transistors
Supported architectures	Conventional and cross bar	Conventional Cellular Neural Networks (CNN)	CNN	Memory based	QCA	Quantum
Cell size (pitch)	100 nm	100 nm	40 nm	Unknown	60 nm	100 nm
Density (#/cm^2)	3×10^{9}	3×10^{9}	6×10^{10}	1×10^{12}	3×10^{10}	3×10^{9}
Switch speed	Unknown	1 THz	1 GHz	Unknown	30 MHz	700 GHz
Circuit speed	30 GHz	30 GHz	1 GHz	Unknown	1 MHz	30 GHz
Switching energy (J)	2×10^{-18}	$>2\times10^{-18}$	1×10^{-18}	1.3×10^{-16}	$>4\times10^{-17}$	2×10^{-18}
Binary throughput (Gbit/ns/cm^2)	86	86	10	Unknown	0.06	86
Operating temperature	RT	RT	20 K	RT	RT or cryogenic	Cryogenic

Source: Selected from 2004 International Technology Roadmap for Semiconductors, http://www.itrs.net/Common/2004Update/2004Update.htm

limits than MOS devices and some are new non charge-based device concepts. In all cases, any new device must provide very low power, be compatible with high levels of integration at very low cost and must provide logic level gain in a three-terminal device. Since, relatively little is known about how such devices could be integrated in large numbers, many of the entries in the table must be taken as best estimates. Only the future will tell if any of these devices will replace MOS devices or find a role in complementing MOS devices.

1.6 Conclusions

It is obvious that Si MOS devices will be pushed to their ultimate limits of scaling. Exactly what technology generation is the ultimate limit for Si devices is somewhat debatable and many experts in the past have predicted the premature demise of scaled Si MOS devices. History has proven that Si MOS device concepts can be pushed much further that most experts envision at any particular point in time. However, there are limits both from a physics point of view and perhaps from an economic point of view. As discussed in previous sections, the golden era of easy scaling has passed and scaling beyond about the 90-nm node becomes increasingly difficult and increasingly requires more new materials and innovative concepts for continued scaling. Scaling may continue to the end of the roadmap (in about 2018) or scaling may simply become too difficult or too costly to continue to the end of the roadmap. In any case, scaling Si MOS devices to the 16-nm node will require significant innovative concepts to achieve the predicted dimensions and device performance.

To achieve the ultimate limits of semiconductor devices the general trends in manufacturing must certainly be toward ever thinner dimensions in all vertical as well as all horizontal dimensions. This in turn of necessity must push manufacturing toward lower thermal budget processing. At the same time from economic reasons, manufacturing is pushed toward larger wafer sizes. The continued downsizing of MOS devices will require that many new materials such as, high-k gate dielectrics and metal gates be introduced into the manufacturing process. New processes, such as selective epitaxy and selective depositions will be required. Device structures will require the control of interface layers, such as the dielectric–silicon interface to atomic dimensions. Such control will in turn require higher levels of automation in the manufacturing process and more in-situ monitoring of the manufacturing process. Although the continued scaling of MOS devices is becoming more difficult, the IC industry appears equal to the task of pushing MOS devices to the ultimate limits at near atomic dimensions.

References

1. Arora, N., *MOSFET Models for VLSI Circuit Simulation*. New York: Springer, 1993.
2. Marcyk, G. "High Performance Non-Planar Tri-gate Transistor Architecture." INTEL publication, ftp://download.intel.com/technology/silicon/Marcyk_tri_gate_0902.pdf (accessed on February, 2007).
3. The National Technology Roadmap for Semiconductors 1997 Edition, published by Semiconductor Industry Association 1997.
4. International Technology Roadmap for Semiconductors, http://www.itrs.net/Links/2004Update/2004Update.htm (accessed on February, 2007).
5. Stern, F., and W. E. Howard. "Properties of Semiconductor Surface Inversion Layers in the Electric Quantum Limit." *Phys. Rev.* 163 (1967): 817.
6. Stern, F. "Self-Consistent Results for n-Type Si Inversion Layers." *Phys. Rev. B* 5 (1972): 4891.
7. Stern, F. "Quantum Properties of Surface Space-Change Layers." *CRL Crit. Rev. Solid State Sci.* 4 (1974): 499.
8. Van Dort, M. J., P. H. Woerlee, and A. J. Walker. "A Simple Model for Quantization Effects in Heavily-Doped Silicon MOSFETs at Inversion Conditions." *Solid State Electron.* 37 (1994): 411.

9. Rios, R., and N. D. Arona. "Determination of Ultra-Thin Gate Oxide Thickness for CMOS Structures Using Quantum Effects." *IEDM Tech. Digest.* (1994): 613.
10. Takagi, S., A. Toriumi, M. Iwase, and H. Tango. "On the Universality of Inversion Layer Mobility in Si MOSFETs: Part I—Effects of Substrate Impurity Concentration." *IEEE Trans. ED* 41 (1994): 2357.
11. Hu, C. "Gate Oxide Scaling Limites and Projections." *IEDM Tech. Digest.* (1996): 319.
12. Taur, Y., D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S-H. Lo, G. A. Sai-Halong. et al. "CMOS Scaling into the Nanometer Regime." *Proc. IEEE* 85 (1977): 486.
13. Momose, H. S., M. Ono, T. Yoshitomi, T. Ohgmo, S. Nakamma, M. Saito, and H. Iwai. "Tunneling Gate Oxide Approach to Ultra-High Current Device in Small Geometry MOSFETs." *IEDM Tech. Digest.* (1994): 593.
14. Kane, E. O. "Theory of Tunneling." *J. Appl. Phys.* 32 (1961): 83.
15. Parker, C., G. Lucovsky, and J. R. Hauser. "Ultrathin Oxide–Nitride Gate Dielectric MOSFETs." *IEEE Trans. ED,* 19 (1998): 106.
16. Yang, H., and G. Lucovsky. "Integration of Ultrathin (1.6–2.0 nm) RPECVD Oxynitride Gate Dielectrics into Dual Poly-Si Gate Submicron CMOSFETs." *IEDM Tech. Digest.* (1999): 245.
17. Groeseneken, G., L. Pantisano, L. A. Ragnarsson, R. Degraeve, M. Houssa, T. Kauerauf, P. Roussel, S. De Gendt, and M. Heyns. "Achievements and Challenges for the Electrical Performance of MOSFETs with High-*k* Gate Dielectrics." *Symp. Phys. Failure Anal. IC* (2004): 147.
18. Gusev, E. P., D. A. Buchanan, E. Cartier, A. Kumar, D. DiMaria, S. Guha, A. Callegari. et al. "Ultrathin High-*k* Gate Stacks for Advanced CMOS Devices." *IEDM Tech. Digest.* (2001): 451.
19. Lo, G. D., D. L. Kwong, K. J. Abbott, and D. Nagarian. "Thickness Dependence of Charge-Trapping Properties in Ultrathin Thermal Oxides Prepared by Rapid Thermal Oxidation." *J. Electrochem. Soc.* 140 (1993): L16.
20. Torimi, A., J. Koga, H. Satake, and A. Ohata. "Performance and Reliability Concerns of Ultra-Thin Gate Oxides MOSFETs." *IEDM Tech. Digest.* (1995): 847.
21. Lu, C-H. , G. M. T. Wong, M. D. Deal, W. Tsai, P. Majhi, C. O. CHus, M. R. Visokay, et al. "Characteristics and Mechanism of Tunable Work Function Gate Electrodes Using a Bilayer Metal Structure on SiO_2 and HfO_2." *IEEE Elect. Dev. Lett.* 26 (2005): 445.
22. Lee, J. H., H. Zhong, Y-S. Suh, G. Heuss, J. Gurganus, B. Chen, and V. Misra. "Tuanble Work Function Dual Metal Gate Technology for Bulk and Non-Bulk CMOS." *IEDM Tech. Digest.* (2002): 359.
23. Lee, J. H., Y-S. Suh, H. Lazar, R. Jha, J. Gurganus, Y. Lin, and V. Misra. "Compaibility of Dual Metal Gate Electrodes with High-*k* Dielectrics for CMOS." *IEDM Tech. Digest.* (2003): 232.
24. Kim, Y. H., C. H. Lee, T. S. Jeon, W. P. Bai, C. H. Choi, S. J. Lee, L. Xinjian, R. Clarks, D. Roberts, and D. L. Kwong. "High Quality CVD TaN Gate Electrode for Sub-100 nm MOS Devices." *IEDM Tech. Digest.* (2001): 667.
25. Datta, S., G. Dewey, M. Doczy, B. S. Doyle, B. Jin, J. Kavalieros, R. Kotlyar, M. Metz, N. Zelick, and R. Chau. "High Mobility Si/SiGe Strained Channel MOS Transistors with HfO_2/TiN Gate Stack." *IEDM Tech. Digest.* (2003): 653.
26. Suh, Y. S., G. Heuss, H. Zhong, S. N. Hong, and V. Misra. "Electrical Characteristics of $TaSi_xN_y$ Gate Electrodes for Dual Gate Si-CMOS Devices." *Digest. 2001 Symp. VLSI Tech.* 47 (2001).
27. Yu, H. Y., J. F. Kang, C. Ren, J. D. Chen, Y. T. Hou, C. Shen, M. F. Li, et al. "Robust High-Quality HfN-HfO_2 Gate Stack for Advanced MOS Device Applications." *Elect. Dev. Lett.* 25 (2004): 70.
28. Anil, K. G., A. Veloso, S. Kubicek, T. Schram, E. Augendre, J. F. deMarneffe, K. Devriendt, et al. "Demonstration of Fully Ni-Silicided Metal Gates on HfO_2 Based High-*k* Gate Dielectrics as a Candidate for Low Power Applications." *Digest. 2004 VLSI Tech.* 190 (2004).
29. Frank, D. J., R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H-S. P. Wong. "Device Scaling Limits of Si MOSFETs and Their Application Dependencies." *Proc. IEEE* 89 (2001): 259.
30. Taur, Y., C. H. Wann, and J. Frank. "25 nm CMOS Design Considerations." *IEDM Tech. Digest.* (1998): 789.
31. Moll, J. L., *Physics of Semiconductors*. New York: McGraw-Hill, 1964.

32. Brews, J. R. "Sensitivity of Subthreshold Current to Profile Variations in Long-Channel MOSFETs." *IEEE Trans. ED* (1996): 2164.
33. Yan, R-H. , A. Ourmazd, and F. Lee. "Scaling the Si MOSFET: From Bilk to SOI to Bulk." *IEEE Trans. ED* 39 (1992): 1704.
34. Suzuki, K., T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto. "Scaling Theory for Double-Gate SOI MOSFET's." *IEEE Trans. ED* 40 (1993): 2326.
35. Tosaka, Y., K. Suzuki, and T. Sugii. "Scaling-Parameter-Dependent Model for Subthreshold Swing S in Double-Gate MOSFET's." *IEEE Elect. Dev. Lett.* 15 (1994): 466.
36. Wong, H-S. , D. J. Frank, Y. Taur, and J. M. C. Stork. "Design and Performance for Sub-0.1 μm Double-Gate SOI MOSFET's." *IEDM Tech. Digest.* (1994): 747.
37. Wann, C. H., K. Noda, T. Tanaka, M. Yoshida, and C. Hu. "A Comparative Study of Advanced MOSFET Concepts." *IEEE Trans. ED* 43 (1996): 1742.
38. Auth, C. P., and D. Plummer. "Scaling Theory for Cylindrical, Fully-Depleted, Surround-Gate MOSFET's." *IEEE Elect. Dev. Lett.* 18 (1997): 74.
39. Pimbley, J. M. "Two-Dimensional Current Flow in the MOSFET Source-Drain." *IEEE Trans. ED,* ED-33 (1986): 986.
40. Ng, K. K., and W. T. Lynch. "Analysis of the Gate-Voltage-Dependent Series Resistance of MOSFETs." *IEEE Trans. ED* ED-33 (1986): 965.
41. Ng, K. K., and W. T. Lynch. "The Impact of Intrinsic Series Resistance on MOSFET Scaling Resistance of MOSFETs Scaling." *IEEE Trans. ED,* ED-34 (1987): 503.
42. Tsui, B-Y. , and M-C. Chen. "Series Resistance of Self-Aligned Silicided Source/Drain Structure." *IEEE Trans. ED* 40 (1993): 197.
43. Ng, K. K., R. J. Bayrens, and S. C. Fang. "The Spreading Resistance of MOSFETs." *IEEE Elect. Dev. Lett.* EDL-6 (1995): 195.
44. Chang, C. Y., Y. K. Fang, and S. M. Sze. "Specific Contact Resistance of Metal-Semiconductor Barriers." *Solid State Electron.* 14 (1971): 54.
45. Chieh, Y-S. , J. P. Krusius, D. Green, and M. Ozturk. "Low-Resistance Bandgap-Engineered W/$Si_{1-x}Ge_x$/Si Contacts." *IEEE Elect. Dev. Lett.* 17 (1996): 360.
46. Gannavaram, S., N. Pesovic, and M. Ozturk. "Low Temperature ($\leq$800°C) Recessed Junction Selective Silicon–Germanium Source/Drain Technoloby for sub-70 nm CMOS." *IEDM Tech. Digest.* (2000): 437.
47. Ozturk, M., J. Liu, H. Mo, and N. Pesovic. "Advanced $Si_{1-x}Gex_x$ Source/Drain and Contact Technologies for Sub-70 nm CMOS." *IEDM Tech. Digest.* (2002): 375.
48. Liu, J., and M. Ozturk. "Nickel Germanosilicide Contacts Formed on Heavily Boron Doped $Si_{1-x}Gex$ Sourec/Drain Junctions for Nanoscale CMOS." *IEEE Trans. ED* 52 (2005): 1535.
49. Nishi, Y. and R. Doering, eds. *Handbook of Semiconductor Manufacturing Technology*, 19. NY: Marcel Decker, 2000.
50. Huang, X., W-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, et al. "Sub 50 nm FinFET: PMOS." *IEDM Tech. Digest.* (1999): 67.
51. Choi, Y-K. , N. Lindert, P. Xuan, S. Tang, D. Ha, E. Anderson, T-J. King, J. Bokor, and C. Hu. "Sub-20 nm CMOS FinFET Technologies." *IEDM Tech. Digest.* (2001): 421.
52. Hisamoto, D., T. Kaga, and E. Takeda. "Impact of the Vertical SOI 'DELTA' Structure on Planar Device Technology." *IEEE Trans. ED* 38 (1991): 1419.
53. Wong, H-S. P. , D. J. Frank, and P. M. Solomon. "Device Design Considerations for Dougle-Gate, Ground-Plane and Single-Gate Ultra-Thin Dual-Gate MOSFETs a the 25 nm Channel Length Generation." *IEDM Tech. Digest.* (1998): 407.
54. Doyle, B., R. Arghavani, D. Barlage, S. Datta, M. Doczy, J. Kavalieros, A. Murthy, and R. Chau. "Transistor Elements for 30 nm Physical Gate Lengths and Beyond." *Intel Tech. J.* 6 (2002): 42.
55. Li, Y., S-M. Yu, C-S. Tang, and T-S. Chao. "Comparison of Quantum Correction Models for Ultrathin Oxide Single- and Double-Gate MOS Structures Under the Inversion Condition." *Proc. 2003 IEEE Conf. Nanotech.* 1 (2003): 36.
56. Welser, J., J. L. Hoyt, S. Takagi, and J. F. Gibbons. "Strain Dependence of the Performance Enhancement in Strained-Si n-MOSFETs." *IEDM Tech. Digest.* 94 (1994): 373.

57. Rim, K., S. Koester, M. Hargrove, J. Chu, P. M. Mooney, J. Ott, T. Kanarsky, et al. "Strained Si NMOSFETs for High Performance CMOS Technology." *Tech. Digest. 2001 VLSI Symp.* (2001): 59.
58. Hoyt, J. L., H. M. Nayfeh, S. Eguchi, I. Aberg, G. Xai, T. Drake, E. A. Fitzerald, and D. A. Antoniadis. "Strained Silicon MOSFET Technology." *IEDM Tech. Digest.* (2001): 23.
59. Rim, K., J. Chu, H. Chen, K. A. Jenkins, T. Kanarsky, K. Lee, A. Mocuta, et al. "Characteristics and Device Design of Sub-100 nm Strained Si N- and PMOSFETs." *Tech. Digest. 2002 VLSI Symp.* (2002): 98.
60. Ieong, M., B. Doris, J. Kenzierski, D. Rim, and M. Yang. "Silicon Device Scaling to the Sub-10 nm Regime." *Science* 306 (2004): 2057.
61. Sze, S. M., *Physics of Semiconductor Devices*. New York: Wiley Interscience, 1981.
62. Thompson, S. E., M. Armstrong, C. Auth, S. Cea, R. Chau, G. Glass, T. Hoffman, et al. "A Logic Nanotechnology Featuring Strained-Silicon." *IEEE Elect. Dev. Lett.* 25 (2004): 191.
63. Thompson, S. E., M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, et al. "A 90-nm Logic Technology Featuring Strained-Silicon." *IEEE Trans. ED* 51 (2004): 1790.

2

Overview of Interconnect—Copper and Low-*k* Integration

Girish A. Dixit
Robert H. Havemann
Novellus Systems, Inc.

2.1 Introduction **2**-1
2.2 Dual Damascene Copper Integration **2**-6
2.3 Copper/Low-*k* Reliability **2**-16
2.4 Conclusion **2**-19
References **2**-20

Over the past decade, integrated circuit scaling and performance needs have driven significant changes in interconnect materials and processes at each successive technology generation. Foremost among these changes has been the transition from aluminum to copper conductors [1–7]—a transition that is virtually complete for logic devices and now underway for memory devices [8,9]. The primary impetus for this ongoing transition has been a need for the improved performance afforded by copper's lower resistivity as compared with aluminum as well as by copper's ability to accommodate higher current densities. The need for improved performance has also driven a concomitant change in the insulator surrounding the conductor, which for logic devices has transitioned from the traditional silicon dioxide dielectric to materials with lower dielectric constant (low-*k*), such as F-doped oxides and C-doped oxides. The simultaneous integration of copper with low-*k* dielectrics presented a significant challenge to the industry, and, while the manufacturing use of copper interconnects has become pervasive, each successive technology generation offers new challenges in terms of meeting density, performance, and reliability requirements. This chapter will provide an overview of copper and low-*k* interconnect integration including process architectures, materials, performance, and reliability issues as well as future scaling challenges and potential technology directions.

2.1 Introduction

The Information Revolution and enabling era of silicon Ultra-Large-Scale-Integration (ULSI) have spawned an ever-increasing level of functional integration on-chip, driving a need for greater circuit density and higher performance. For classical transistor scaling, device performance improves as the gate length and the gate dielectric thickness are scaled. Only recently have new materials, such as high-*k* gate dielectrics and metal gates been considered as essential for continued transistor scaling.

In contrast, as the chip wiring (interconnect) is scaled, performance degrades as both resistance and current density increase due to the smaller cross-sectional area of the scaled conductor. The introduction of copper metallization served as an enabler for continued interconnect scaling due to its lower resistivity

(~1.8 μ-ohm-cm) as compared with traditional AlCu metallization (~3.3 μ-ohm-cm) as well as its ability to accommodate higher current density [10,11].

An additional consequence of scaling is an increase in sidewall capacitance as conductors are placed in closer proximity to one another. While metal thickness can be reduced to mitigate the increase in sidewall capacitance, the consequences of this approach are increased resistance and current density. Alternative circuit design solutions, such as increasing conductor spacing and/or adding extra levels of interconnect with relaxed design rules, have the drawbacks of reduced density and increased cost. The introduction of low-*k* dielectrics provided a materials solution that mitigated sidewall capacitance [12–18] and provided more latitude in the co-optimization of process architecture and circuit design.

Previous analyses [19–25] have highlighted the interconnect performance issues that are incurred as integrated circuit design rules continue to scale. As illustrated in the International Technology Roadmap for Semiconductors (ITRS) (see Figure 2.1), a chief concern is the increasing latency or Resistance-Capacitance (as in RC delay of inverter circuit) (RC) delay of global wiring [26,27]. Since local and intermediate interconnects tend to scale in length, latency is dominated by global interconnects connecting large functional logic blocks, as shown in Figure 2.2. Future increase in microprocessor chip size predicted by the ITRS [27] bring heightened concern, since interconnect latency is proportional to the square of the length. While design solutions such as the use of repeaters (as shown in Figure 2.1) or reverse scaling may mitigate latency in the near term, these approaches typically result in larger chip size and/or more levels of interconnect, leading to higher product cost.

For local and intermediate wiring levels, crosstalk is an additional interconnect performance issue that must be considered. Signal crosstalk is given by the ratio of line-to-line (sidewall) capacitance to total capacitance as shown in Figure 2.3. As transistor operating voltage continues to scale downward, interconnect crosstalk and noise levels must be reduced to avoid spurious transistor turn-on. Since crosstalk is dominated by interconnect sidewall capacitance (as is overall capacitance for minimum feature size as shown in Figure 2.3), process-related solutions, such as the use of thinner metallization

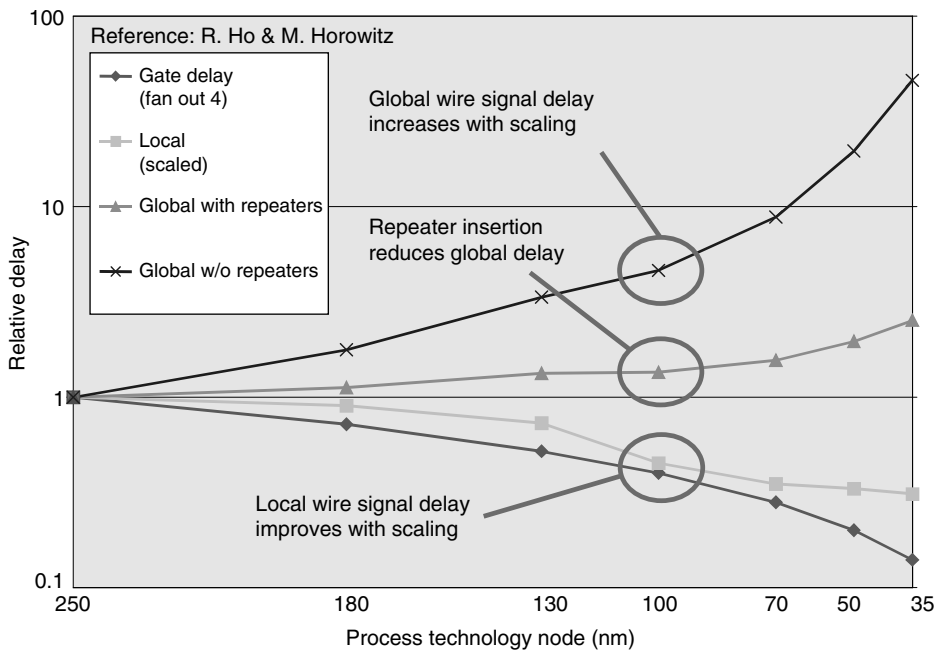


FIGURE 2.1 Relative delay for logic gate Fan Out (as in Fan Out of inverter circuit) (FO = 4) vs. both local and global interconnects as a function of technology node. (From Deodhar, V. V. and Davis, J. A., *Tech. Dig. Int. Symp. Circuits Syst.*, V-349–V-352, 2003.)

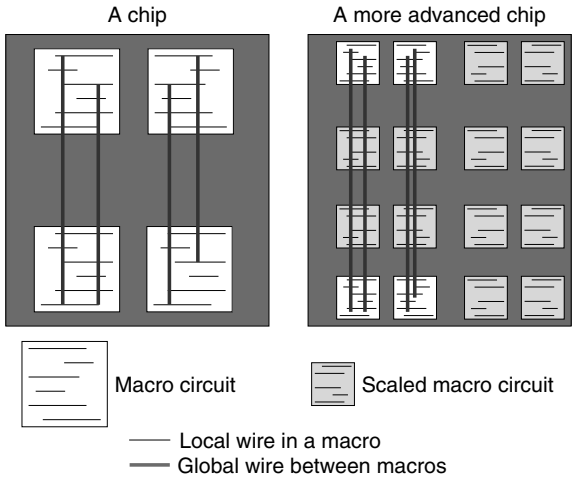


FIGURE 2.2 Example scaling of global vs. local interconnects: length of global interconnects does not scale relative to chip size, whereas the length of local interconnects is reduced by the scaling factor.

and/or low-*k* dielectrics must be implemented to enable continued scaling. This is yet another reason why copper and low-*k* dielectrics have become an essential part of the Integrated Circuit (IC) scaling engine.

As operating frequency continues to increase power dissipation in the interconnect system, which is proportional to both switching frequency and capacitance, has become a significant portion of the overall power dissipated in the chip as shown in Table 2.1 and discussed in several recent papers [28–37]. Thus, the need to limit interconnect power dissipation provides yet another impetus for capacitance reduction in addition to latency concerns. Typical high performance designs utilize a hierarchical or "reverse scaling" metallization scheme (Figure 2.4), where widely spaced "fat wires" are used on an upper global interconnect and power levels to minimize RC delay and voltage drop. Maintaining power distribution at constant voltage through equipotential wires to all V_{dd} bias points requires increasingly lower resistance global wires as operating voltage continues to scale and switching frequencies increase. This need has been partially addressed by the introduction of ball-grid-array packaging technology [38–40] that

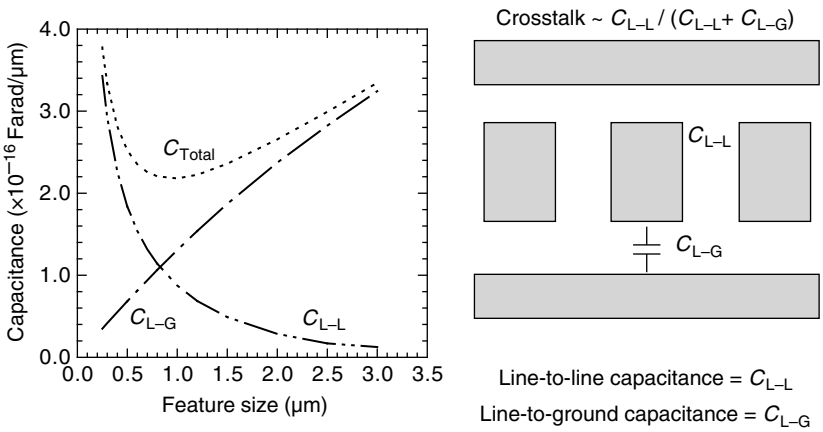


FIGURE 2.3 Simulation of interconnect capacitance as a function of feature size assuming fixed metal height and equal line/space. Line-to-line capacitance dominates as feature size decreases. Interconnect crosstalk is given by line-to-line capacitance divided by the total capacitance.

TABLE 2.1 Selected Overall, Technology Characteristics Based on the 2005 International Technology Roadmap for Semiconductors

Year	2005	2007	2010	2013	2016	2019
MPU (Microprocessor Unit) ½ pitch	90	68	45	32	22	16
MPU patterned gate length (nm)	54	48	30	21	15	11
DRAM ½ pitch (nm)	80	65	45	32	22	16
MPU: frequency of on-chip clock for high performance (MHz)	5204	9285	15,079	22,980	39,683	62,443
MPU: frequency of chip to board for high performance (MHz)	3125	4883	9536	18,625	36,378	71,051
High volume-MPU (cost-performance) chip size at production (mm^2)	111	140	140	140	140	140
High volume-MPU (cost-performance) Mtransistors/cm^2	174	276	552	1104	2209	4,417
DRAM memory chip size at production (mm^2)	88	110	93	93	93	93
DRAM memory chip Gbits/cm^2	1.22	1.94	4.62	9.23	18.46	36.93
Maximum power w/high performance heat sink (watts)	167	189	198	198	198	198
Maximum power (watts)—battery (Hand-held)	2.8	3.0	3.0	3.0	3.0	3.0
Minimum logic V_{dd}—(maximum performance) volts	1.1	1.1	1.0	0.9	0.8	0.7
Minimum logic V_{dd}—(lowest power) volts (battery power)	0.9	0.8	0.7	0.60	0.50	0.5

Source: From *International Technology Roadmap for Semiconductors*, published by the Semiconductor Industry Association, 2005.

distributes individual power feeds across the chip. However, new packaging technologies will undoubtedly be required to alleviate the increasing level of power dissipation generated on-chip.

Over the past decade, the aforementioned device scaling and performance needs have driven dramatic changes in interconnect materials and processes at each successive technology generation. While the motivation for moving from aluminum to copper metalization and from oxide to low-*k* dielectrics is clear, significant material and process innovation has been and will continue to be required to meet the interconnect goals set forth in the ITRS. A summary of key interconnect requirements from the ITRS [26] is shown in Table 2.2. The smaller feature sizes and higher aspect ratios projected for copper dual damascene structures necessitate thinner and more conformal metal barriers to prevent copper diffusion into surrounding dielectrics. While advanced physical vapor deposition (PVD) barrier/seed deposition technologies have proven to be extendable to at least the 45 nm technology node new metal deposition techniques, such as atomic layer deposition (ALD), will ultimately be required to achieve ultra-thin

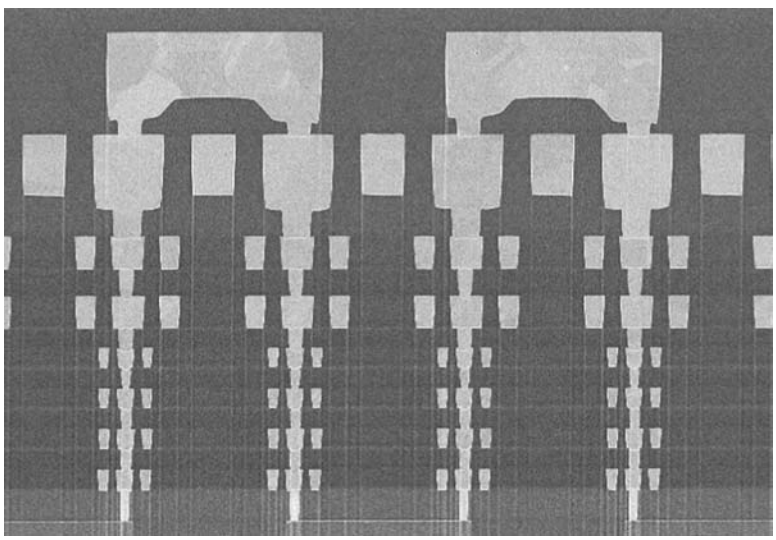


FIGURE 2.4 Example of hierarchical interconnect architecture used in 90 nm node digital signal processor (DSP) (Courtesy of Texas Instruments).

TABLE 2.2 Selected Interconnect Technology Projections from the 2005 International Technology Roadmap for Semiconductors

YEAR	*2005*	*2007*	*2010*	*2013*	*2016*	*2019*
MPU ½ pitch	90	68	45	32	22	16
MPU gate length (nm)	32	25	18	13	9	6
Number of metal levels	11	11	12	13	13	14
Number of optional levels – ground planes/capacitors	4	4	4	4	4	4
Jmax (A)/cm^2 – intermediate wire (at 105 C)	8.91×10^5	2.08×10^6	5.15×10^6	8.08×10^6	1.47×10^7	2.23×10^7
Metal 1 (Cu) wiring pitch (nm)	180	136	90	64	44	32
Metal 1 A/R	1.7	1.7	1.8	1.9	2	2
Metal 1 barrier/cladding thickness (nm)	6.5	4.8	3.3	2.4	1.7	1.2
Cu thinning at minimum Metal 1 pitch due to erosion (nm), 10% x height, 50% areal density, 500 μm square array	15	12	8	6	4	3
Metal 1 effective resistivity (assumes conformal barrier and includes electron scattering effects) (μΩ-cm)	3.15	3.47	4.08	4.83	6.01	7.34
Metal 1 RC delay (ps) over 1mm with effective resistivity	440	767	1792	3451	8040	15853
Intermediate wiring pitch (nm)	200	140	90	64	44	32
Intermediate wiring dual damascene A/R (Cu wire/via)	1.7/1.5	1.8/1.6	1.8/1.6	1.9/1.7	2.0/1.8	2.0/1.8
Intermediate wiring barrier/cladding thickness (nm)	7.3	5.2	3.3	2.4	1.7	1.2
Cu thinning at minimum intermediate pitch due to erosion (nm), 10% x height, 50% areal density, 500 μm square array	17	13	8	6	4	3
Intermediate metal effective resistivity (assumes conformal barrier and includes electron scattering effects) (μΩ-cm)	3.07	3.43	4.08	4.83	6.01	7.34
Intermediate wiring RC delay (ps) over 1mm calculated using effective resistivity shown above	355	682	1825	3504	8147	16059
Minimum global wiring pitch (nm)	300	210	135	96	66	48
Global wiring dual damascene A/R (Cu wire/via)	2.2/2.0	2.3/2.1	2.4/2.2	2.5/2.3	2.6/2.4	2.8/2.5
Global wiring barrier/cladding thickness (nm)	7.3	5.2	3.3	2.4	1.7	1.2
Cu thinning of global wiring due to dishing (nm), 100 μm wide feature	24	19	14	10	8	6
Global metal effective resistivity (assumes conformal barrier and includes electron scattering effects) (μΩ-cm)	2.53	2.73	3.10	3.52	4.20	4.93
Global wiring RC delay (ps) over 1mm calculated using effective resistivity shown above	111	209	523	977	2210	4064
Interlevel metal insulator-effective dielectric constant (k)	3.1– 3.4	2.7 – 3.0	2.5 – 2.8	2.1 – 2.4	1.9 – 2.2	1.6 – 1.9
Minimum expected bulk dielectric constant (k)	≤2.7	≤2.4	2.2	2.0	1.8	1.6

Manufacturing solutions exist
Manufacturable solutions are known
Manufacturable solutions are not known

Source: From *International Technology Roadmap for Semiconductors*, published by the Semiconductor Industry Association, 2005.

barriers. Likewise, while improvements in copper electroplating technology have enabled extendibility through multiple generations, new chemistries and techniques must be developed to accelerate bottom-up fill and plate on high resistivity seeds, as feature size continues to decrease.

Paralleling the development of advanced copper metallization techniques is an equally important effort focused on lower *k* dielectric materials. The integration of new low-*k* dielectrics brings numerous reliability concerns that include thermally- or mechanically-induced cracking or adhesion loss, poor mechanical strength, moisture absorption, lower dielectric breakdown voltage/time dependent dielectric breakdown (TDDB), texture effects and poor thermal conductivity. The reduced mechanical strength of porous low-*k* dielectrics is of particular concern in both processing (especially during chemical–mechanical polishing (CMP)) and packaging. Thus, mechanical properties, such as hardness, modulus, cohesive strength, cracking limit, and crack propagation velocity have been key metrics for ongoing materials development. For so-called "porous" ultra-low-*k* (ULK) materials, co-optimization of deposition chemistry with porogen removal and anneal has yielded dielectric materials that are compatible with advanced manufacturing requirements [41–48]. However, while current progress is encouraging, past history underscores the difficulty of introducing new low-*k* materials into production and much work remains to be done.

In general, a key integration challenge for ULK materials and for scaling copper damascene is the extendibility of CMP techniques. In addition to accommodating more mechanically fragile and chemically hydrophobic dielectric materials, future CMP processes must also contend with tighter dishing and erosion specifications. The decrease in metal thickness dictated by scaling means that to maintain design tolerances for resistance and capacitance, tighter control of vertical dimensions is required in the damascene processes (see Table 2.2). Achieving these goals will likely require co-optimization of both electroplating and planarization techniques as well as enhanced in situ process control.

While significant copper and low-*k* process integration challenges lie ahead, several generations of copper/low-*k* production provide a solid foundation for moving forward. The next section will describe some of the key process integration challenges and solutions that have contributed to the successful implementation of copper/low-*k* interconnects in high volume production.

2.2 Dual Damascene Copper Integration

Subtractive etch, the approach used in fabricating aluminum-based interconnects is inapplicable in the fabrication of copper-based interconnects, due to the lack of volatility of copper–halide complexes at moderate temperatures. As a result, copper interconnect fabrication requires a damascene approach whereby the metallization is inlaid into interconnect geometries which are pattern-transferred into the

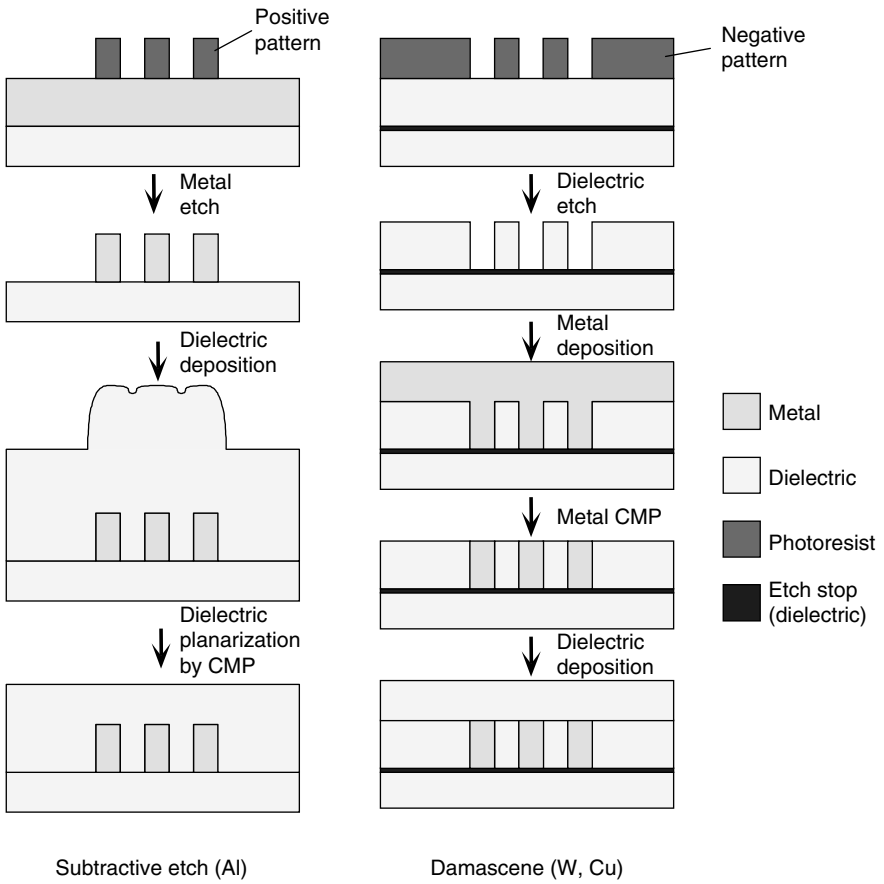
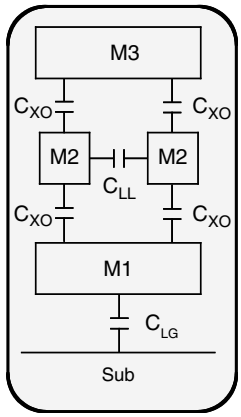


FIGURE 2.5 Comparative process flows associated with the fabrication of interconnect structures using subtractive and damascene technologies.

TABLE 2.3 Key Process and Performance Differences for Subtractive vs. Damascene Interconnect Technologies

Key Differences—Subtractive vs. Damascene		
Subtractive		Dual Damascene
Interconnect resistance variance depends mostly on line width variance (driven by litho/etch bias, wet cleans)		Interconnect resistance variance depends on line width/depth variance (driven by litho/etch bias, wet cleans, etch depth uniformity)
Resistance variance mostly constant with increasing width and thickness		Resistence variance may change with increasing width and thickness
C_{LL} variance depends on line width variance		C_{LL} variance depends on line width/depth variance
C_{XO} variance depends on line width/depth variance		C_{XO} variance depends on line width/depth variance
Via depth fixed for each level of interconnect		Via depth function of layout(width of overlying line)
Self limiting oxidation of Al		Cu oxidation no self limiting
Chemical–mechanical polishing (CMP) limited to single material		Chemical–mechanical polishing (CMP) of composite structure

dielectric of interest. A flow comparison of the subtractive and damascene sequences is illustrated in Figure 2.5.

A dual damascene process also offers lower fabrication cost due to the limited use of chemical–mechanical planarization processes compared to the multiple uses of this unit process in the subtractive etch fabrication of interconnects. In addition, low via resistances are achieved through the reduction of the number of high resistivity interfaces in the interconnect structure. However, the dielectric etches and metal fill processes face higher aspect ratios due to the dual damascene structure. Table 2.3 lists key differences between the two approaches.

Dual damascene copper interconnects may be fabricated using two primary schemes; via first scheme or trench first scheme as outlined in Figure 2.6. Continued scaling of interconnect geometries also requires integrating low permittivity materials into the copper interconnect structure. The chemically amplified materials used in the photolithographic steps of pattern transfer exhibit increased sensitivity to the impurities in the chemical vapor deposited low permittivity dielectric materials. Interactions between impurities (N, H, and combinations of these) in the low permittivity dielectric films and the UV lithography resist lead to a loss of sensitivity of the photoactive compounds in the pattern definition layers. The interaction between the amine groups and photoactive compounds in the resist may lead to

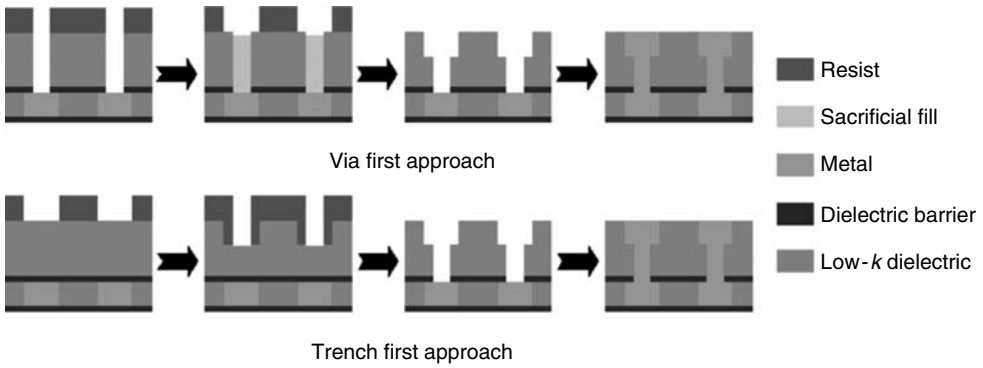


FIGURE 2.6 Comparison of via first vs. trench first dual damascene approaches.

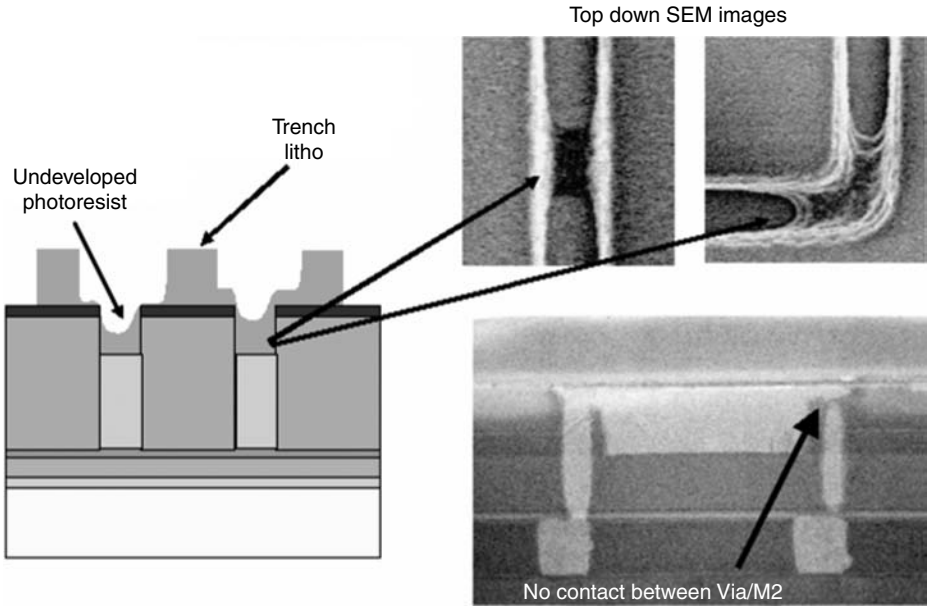


FIGURE 2.7 The schematic representation of resist poisoning (left), scanning electron microscopy (SEM) top down, and cross-section micrographs illustrating the pattern disruption due to undeveloped resist. (From Dixit, G., *Proceedings of the International Reliability Physics Symposium*, Tutorial Notes, 2004.)

undeveloped resist thus preventing the formation of all the required features in a multilevel interconnect structure. Figure 2.7 shows the phenomenon of resist poisoning occurring in the trench pattern step of a via first dual damascene scheme [49]. Various modifications to the sequence and details of process steps used in the fabrication sequence may be employed to overcome the risk of resist poisoning.

Figure 2.8 outlines the various schemes used in dual damascene fabrication [50,51]. In the self-aligned approach shown in Figure 2.8a, the via level dielectric, or interlayer dielectric (ILD) and an etch stop layer (typically silicon nitride or silicon carbide for inorganic ILDs and oxide for organic ILDs) are sequentially deposited, followed by pattern and etch of via into the etch stop layer. The dielectric for the trench is then deposited onto the patterned etch stop layer. The trench features are delineated into this dielectric and the trench etch is extended to complete transferring the via pattern from the etch stop layer into the interlayer dielectric. The etch stop layer defines the trench height, while maintaining a vertical profile of the via sidewall. The etch stop layer is removed from the bottom of the trench during the final etch step, which simultaneously clears the dielectric barrier from the bottom of the via. The chief advantage of the buried via approach is that all patterning is done on planar surfaces; major disadvantages include the need for an etch stop layer (which increases sidewall capacitance), the need for high etch selectivity to the etch stop layer and susceptibility to partial via definition if trench and via are misaligned. Partial vias present a potential reliability issue and, thus, this integration scheme should be avoided unless ample alignment tolerance is provided in the product design.

Via first approach for dual damascene has been a workhorse for the industry. In this scheme, the entire dielectric stack (including intervening etch stop layer if desired) for a given interconnect level is deposited prior to pattern definition. The vias are then patterned and etched down to the etch stop layer as shown in Figure 2.8b. The high aspect ratio vias are filled with a sacrificial inorganic material to protect the underlying etch stop layer during the trench etch. The sacrificial fill material also assists the trench patterning process by limiting the variation in resist thickness over via. The etch rate of the sacrificial fill material is required to be similar or slightly higher than the etch rate of the dielectric during the trench etch. Bottom anti-reflective layers (BARC) are frequently used as sacrificial fills in via first dual damascene.

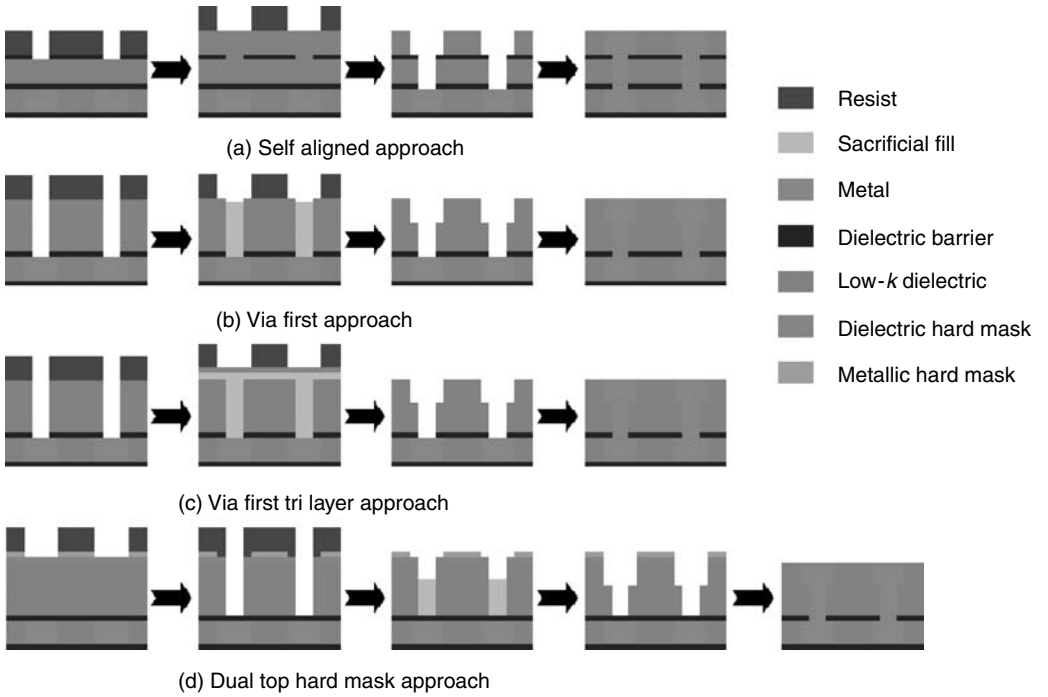


FIGURE 2.8 Schematic flows of improvised via first and trench first approaches to overcome challenges in the pattern transfer process for low-*k* porous dielectrics.

In the trench first approach, via is patterned on the etched trench, which may present significant topography. In case of misalignment, partial vias can only be avoided by extended overetch to ensure that the full ILD/inter-metal dielectric (IMD) thickness is cleared, which taxes etch selectivity to the via etch stop layer (not protected by a BARC as in the case of via first approach).

Via first tri-layer approach (Figure 2.8c) utilizes a sacrificial dielectric film, such as undoped silicon dioxide deposited on top of the sacrificial via fill material, to improve the fidelity of the pattern transfer process. A low temperature dielectric deposition process is preferred to ensure compatibility with the sacrificial fill materials, whose glass transition temperatures reside below those of conventional dielectric deposition processes. The dielectric film serves as a barrier to prevent interaction between the resist and contaminants from the underlying low dielectric constant materials, thus enhancing the robustness towards resist poisoning. The low etch rate of the silicon dioxide during the low-*k* etch process and further separation between the low-*k* dielectric and the UV resist offers advantages in controlling the sidewall roughness of the resultant features.

Figure 2.8d depicts one of the trench first dual damascene sequences with two hard mask layers [52,53]. The sacrificial hard mask stack is comprised of a dielectric/metal nitride bilayer. Materials such as titanium nitride may be used as a metal hard mask. The optical transparency of this material is a key requirement to ensure alignment between the successive pattern steps and the transparency requirements limit the useable thicknesses of these films. The process steps typically involve pattern and etch of the trench features into the metal hard mask followed by via pattern. Following via etch; a sacrificial fill of via may be utilized prior to completing the trench etch with the metal nitride mask. Typical post-etch resist removal processes include an oxidizing plasma which in turn may also cause oxidation and loss of carbon from the sidewalls of the trench features. The oxidation damage to low-*k* materials is undesirable as it leads to an increase in the dielectric permittivity. Due to the absence of resist following the trench etch, an oxidizing plasma clean is unnecessary and the post-etch clean may be accomplished through the use of

TABLE 2.4 Process and Integration Challenges for Different Dual Damascene Schemes

Damascene Schemes	
Approach	Challenges
Self-aligned	Risk of forming partial vias, worse with scaling
	Additional *k* impact due to middle etch stop layer
	Has not been reported in any practical use
Trench first	Difficult to clear resist puddle in trench (depth of focus)
	Risk of misalignment leading to partial vias
	Alignment issue aggravated with scaling
	Resist poisoning
Via first	Fencing (different etch rates of dielectric/sacrificial via fill)
	Resist poisoning and lithography rework
	Low-*k* damage in post etch cleaning
Via first tri layer	Trench etch process requires further optimization to control profiles and limit fencing
	Sacrificial dielectric compatibility with sacrificial fill layer
Dual top hard mask	Multiple materials in stack (etch selectivity management)
	Significant etch process challenge, may require multiple tools
	Top hard mask compatibility with Cu Chemical–mechanical polishing (CMP)

solvents, thus eliminating the risk of oxidizing the low-*k* material. This approach involves a metal etch step to define the trench into the metal nitride and this may require additional process equipment compared to the approaches without metal hard masks. Table 2.4 summarizes the challenges in the various dual damascene schemes.

Integrating low permittivity materials with porosity, high carbon content, and marginal mechanical properties present challenges in the areas of interface engineering, pattern transfer, and metallization. Figure 2.9 summarizes various issues that may be faced during the dual damascene integration of copper with low dielectric constant materials. In order to lower the effective interconnect capacitance between successive generations of integrated circuits, it is necessary to reduce the dielectric constant of the bulk dielectric as well as the dielectric etch stop layer. As a result, the commonly used etch stop layer such as silicon nitride is replaced with nitrogen doped silicon carbide. The presence of carbon in the etch stop

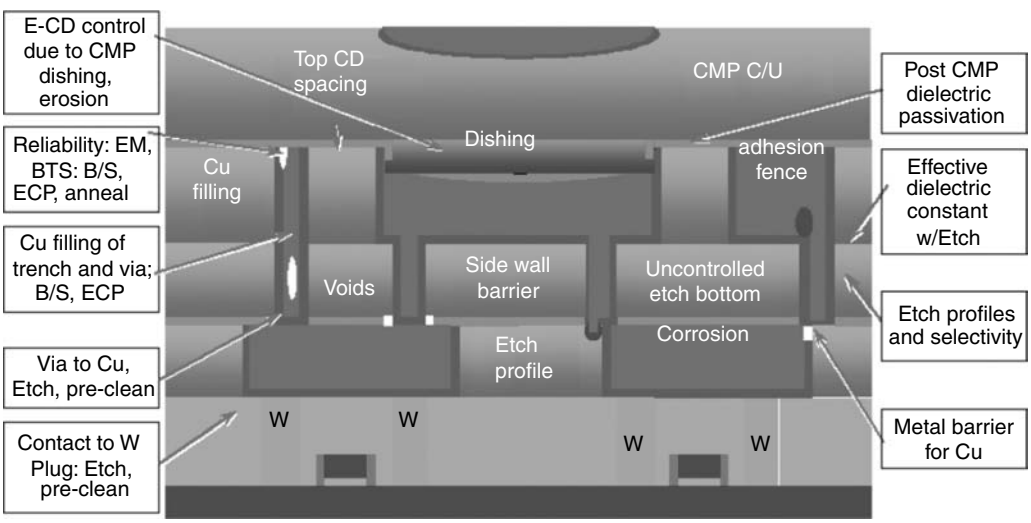


FIGURE 2.9 Schematic representation of issues faced in the integration of Cu/low-*k* dielectric interconnect structures. (From Dixit, G., *Proceedings of the International Reliability Physics Symposium*, Tutorial Notes, 2004.)

layer as well as in the bulk dielectric leads to a concern for interface delamination, as the carbon in both these films may segregate preferentially to the surfaces [54,55]. Innovative pre- and post-dielectric deposition treatments are utilized to denude the surfaces of the excess carbon so as to promote the strong interfacial adhesion between the various films in the dielectric stack. The local changes in the film concentrations may in turn present challenges to the dielectric etch and post-etch cleaning processes as the etch rates at the interfaces may differ from the etch rates of the bulk materials.

Plasma etch of the multilayer dielectric stack challenges the etch unit processes in maintaining the etch selectivity between the different layers, while simultaneously producing acceptable profiles of the resultant via and trench features [56,57]. Intervening etch stop layers in via/trench architecture degrade the effective capacitance of the structure and such layers are undesirable. Eliminating the intermediate etch stop may result in features with non-optimal shapes, such as facets, micro-trenches, or fences as shown in Figure 2.10. Tailoring the etch unit process to meet the varying requirements of etch rate and selectivity between the different materials, with minimal loading effects, is of primary importance in designing out the intermediate etch stop layers from the dual damascene structure.

The impact of resist removal processes on the dielectric properties of low dielectric constant materials needs careful evaluation [58–61]. The chemical plasmas used for stripping the residual resist and by-products of etch adversely affect the carbon concentration of the low-*k* materials. The porous structure of the low-*k* material is then susceptible to adsorb moisture and may exhibit a large increase in permittivity. Figure 2.11 shows the impact on the final trench structure produced after etch, ash, and solvent clean. Undesirable undercut of the sidewall is noted with the case of an unoptimized resist removal process. Figure 2.12 shows the lateral carbon concentration profile on the trench sidewall for structures exposed to different resist removal plasmas. A reduction in the carbon concentration close to the sidewall is seen for both types of processes and the width of the zone with variable carbon content differs significantly between the two processes. These physical changes to the low-*k* dielectric material lead to alteration of the electrical properties as seen in Figure 2.13 where the normalized product of resistance and capacitance of an interconnect structure is plotted for different resist removal processes. The choice of an optimum etch in the dual damascene scheme is thus of high importance to achieve effective electrical performance.

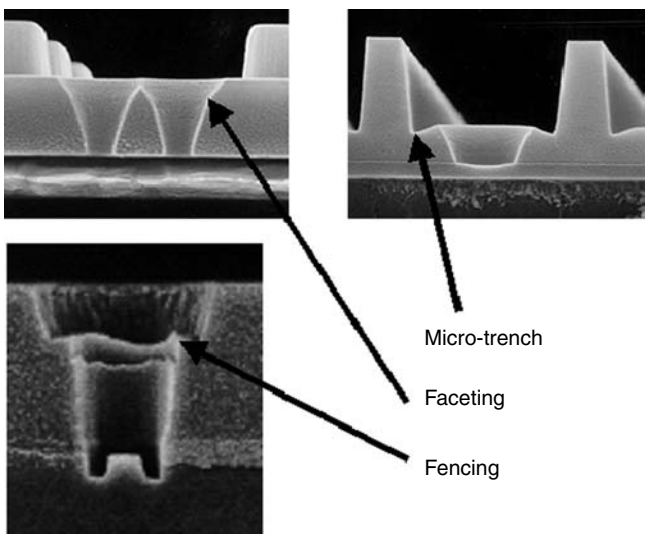


FIGURE 2.10 SEM cross-section images illustrating undesirable results, such as micro-trenching, faceting, and fencing encountered due to non-optimized pattern transfer. (From Dixit, G., *Proceedings of the International Reliability Physics Symposium*, Tutorial Notes, 2004.)

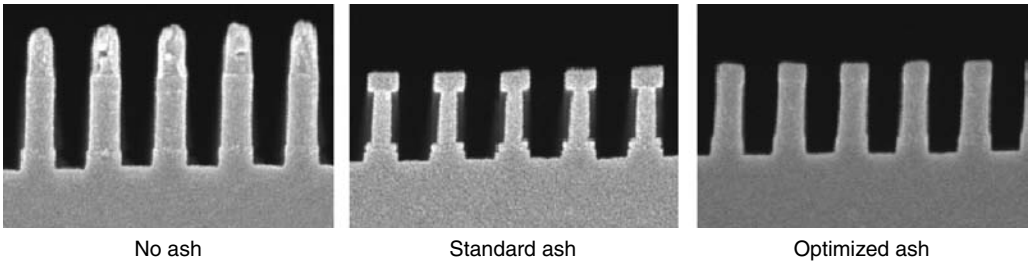


FIGURE 2.11 SEM cross-sections of trenches etched into a porous low-*k* dielectric. The post-etch ash induced damage to the dielectric is noted in the form of an undercut of the low-*k* material by the buffered hydrofluoric acid (HF) decoration and results in unwanted line width gain. (From Dixit, G., *Proceedings of the International Reliability Physics Symposium*, Tutorial Notes, 2004.)

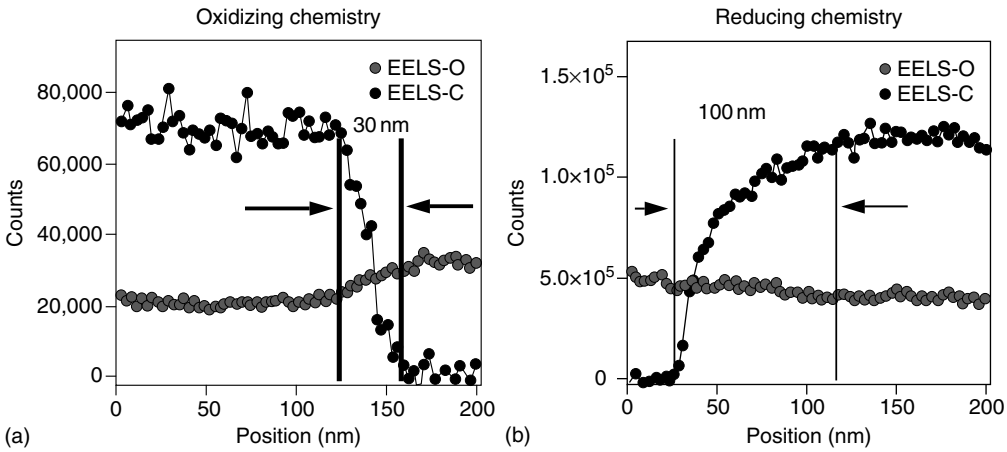


FIGURE 2.12 Transmission electron microscope (TEM) electron energy loss spectroscopy (EELS) profiles of carbon and oxygen concentrations across the dielectric between trenches. Varying widths of the carbon loss zone are noted with different post-etch ash chemistries. (From Dalton, T. J., Fuller, N., Tweedie, C., Dunn, D., Labelle, C., Gates, S., Colburn, M., et al., *Tech. Dig. IEEE Int. Interconnect Tech. Conf.*, 154–56, 2004.)

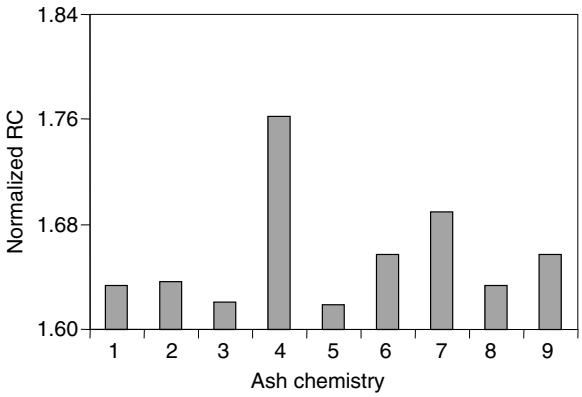


FIGURE 2.13 A plot showing a normalized RC product, measured on an interdigitated metal comb structure, as a function of different post-etch ash processes. Damage to the dielectric leads to higher capacitance and higher RC product. (From Dixit, G., *Proceedings of the International Reliability Physics Symposium*, Tutorial Notes, 2004.)

In order to address the downstream process compatibility, dual damascene features fabricated in low-*k* dielectric materials may contain sacrificial layers, such as dielectric or metal hard masks, which remain prior to metallization and are removed during the chemical–mechanical polish of copper. These sacrificial layers contribute to increase the aspect ratios of the damascene feature and additional stress on the dielectric etches and copper fill processes. The liner/barrier processes for contact and via schemes typically includes an inert ion based sputter pre-clean that ensures consistent ohmic contact between the under/overlying metallic layers. The sputter pre-clean redistributes the material removed from the interface of interest and the sputtered species are deposited onto the sidewall of the features (Figure 2.14). In the case of copper low-*k* interconnects, the redeposited film presents an additional challenge, as this material may then diffuse into the porous dielectric, thus degrading the intra-metal isolation and increasing the risk of copper contamination of the entire structure. The outdiffusion of the redeposited material into the dielectric and the resultant free volume within the interconnect feature also contributes to increased risk in stress migration of interconnects.

The high aspect ratios due to the presence of sacrificial layers in the dual damascene flow challenge the capability of the physical vapor deposition process in providing continuous conformal deposition. The ease of integration between the physical vapor deposited layers and subsequent electroplating of copper has led to the need to extend the physical vapor deposition process for copper barrier and seed layers. In order to overcome the line-of-sight limitations of physical vapor deposition, innovative approaches such as resputtering of the barrier layer are utilized to improve the coverage of the barrier layers on the sidewalls of the vias and trenches. The resputter in copper barrier deposition may also be used to eliminate the inert ion sputter pre-clean process. In the barrier first approach [62], the copper barrier resputter step is tuned to achieve a controlled penetration of the via into the underlying metal. Due to the varying aspect ratios of via in the dual damascene architecture, the bottom coverage of barrier may vary significantly in vias within different trench geometries. As a result, in achieving a controlled recess of the via, the dielectric in certain trench geometries may be exposed during the resputter step, thereby leading

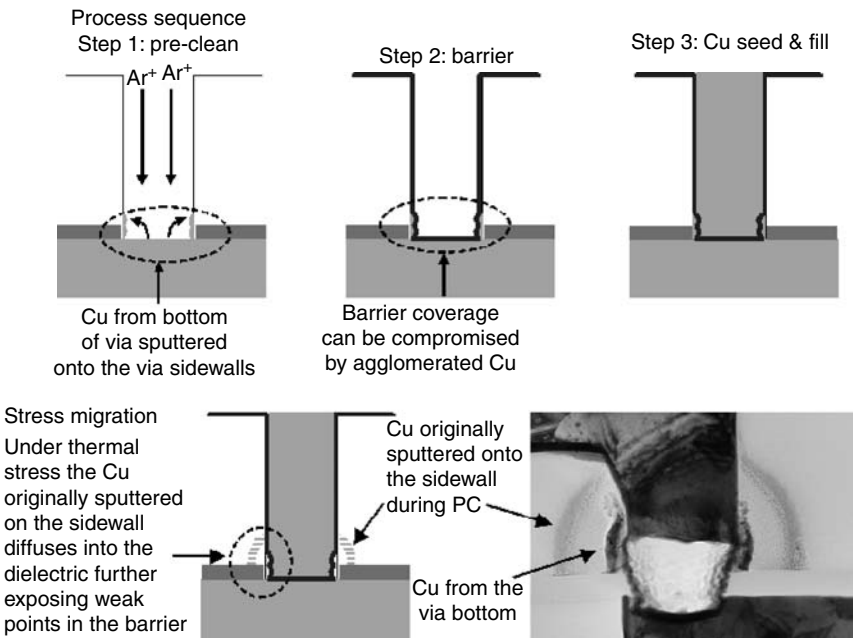

Process sequence
Step 1: pre-clean
Ar+ Ar+
Cu from bottom of via sputtered onto the via sidewalls
Step 2: barrier
Barrier coverage can be compromised by agglomerated Cu
Step 3: Cu seed & fill
Stress migration
Under thermal stress the Cu originally sputtered on the sidewall diffuses into the dielectric further exposing weak points in the barrier
Cu originally sputtered onto the sidewall during PC
Cu from the via bottom

FIGURE 2.14 Schematic representations of process sequence of pre-clean in the Cu barrier/seed deposition and its impact on failure of via (TEM cross-section) subjected to thermal stress. (From Dixit, G., *Proceedings of the International Reliability Physics Symposium*, Tutorial Notes, 2004.)

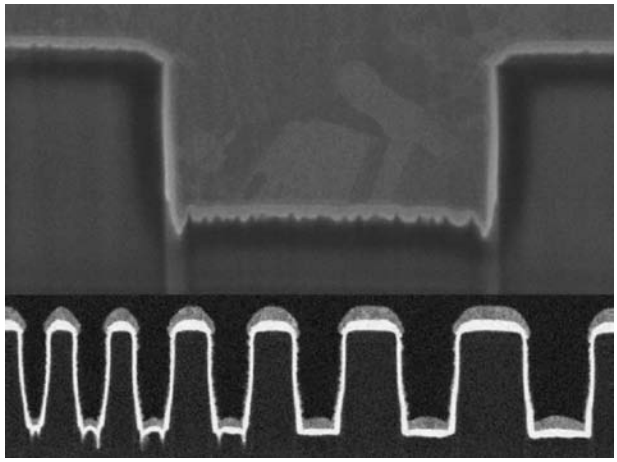


FIGURE 2.15 SEM cross-sections illustrating the damage that may be caused to the trench dielectric due to un-optimized resputtering in the copper barrier deposition process.

to micro trenching or trench bottom roughening as seen in Figure 2.15. The rough trench bottom poses a reliability risk due to the possibility of micro voids between the barrier and copper seed interface. Precise control of the barrier deposition parameters to achieve differentiated barrier thicknesses in the bottom of trench and via features is necessary to eliminate damage to the dielectric and roughening during the resputtering process.

The composition of the Ta(N) layer as well as the crystallographic texture of the Ta deposited on top of the Ta(N) play a key role in determining the continuity of the copper seed deposited on top of the barrier [63]. A continuous seed with good adhesion is necessary to ensure complete filling of the electroplated copper into the dual damascene features. As the interconnect geometries scale into the sub-100-nm regime, the thickness of the copper seed deposited in these features also needs to be scaled down to prevent pinch-off prior to the electroplating step. Electroplating on thin copper seeds with relatively high sheet resistance is challenged by the terminal effect in the plating cell, whereby the voltage drop across the wafer diameter induces a large non-uniformity in the electroplated film thickness and may also lead to voiding within the features. Continued development in barrier/seed deposition and electroplating of copper is required to extend these processes to future generations of interconnect.

The resistivity of copper in sub-50 nm features is significantly higher compared to the resistivity in wider lines (Figure 2.16) [64,65]. The relatively high volume of these narrow lines occupied by the highly resistive Ta(N)/Ta barriers, smaller grain size of the copper in narrow features compared to wider features as well as the sidewall roughness of the damascene trenches all contribute to the increased resistivity. Thin high efficiency barriers, such as atomic layer deposited Ta(N) and integration of these materials with thin low resistivity seed layers for electroplating is necessary to ensure the extendibility of copper metallization to decanano scale interconnect features.

A number of difficulties are encountered in the chemical–mechanical planarization processes of copper low-*k* dielectrics. Carbon doped low dielectric constant materials exhibit lower hardness and modulus. While the dual damascene pattern transfer may be accomplished through the use of suitable hard mask layers, in order to realize the maximum benefit in overall capacitance of the interconnect sacrificial masking layers need to be removed during the chemical–mechanical planarization processes. The exposure of the low-*k* materials to the polishing environment presents hurdles in maintaining acceptable dishing and erosion of the copper low-*k* structures during chemical–mechanical planarization processes.

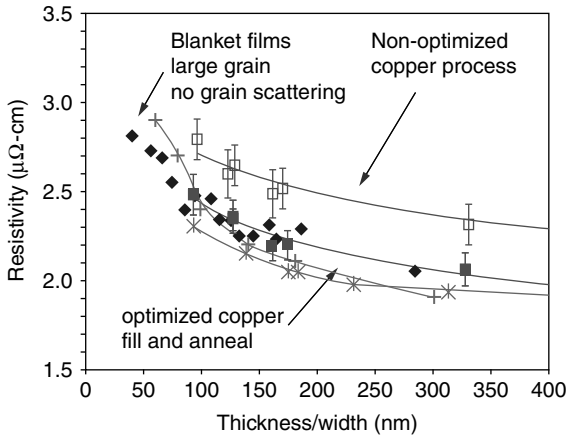


FIGURE 2.16 Measured copper resistivity as a function of line width. Resistivity increases at smaller line widths due to electron scattering from grain boundaries and sidewalls. Optimization of plating chemistry and post-plate anneal for large grain growth reduces grain boundary scattering.

Precise control of the shear force applied during the polishing process is necessary to achieve good process performance as well as to minimize mechanical damage to the low-k materials [66–73]. The hydrophobic nature of porous low-k materials with inferior wetting characteristics requires the implementation of advanced vapor drying techniques to minimize defects, such as residues of the polishing slurry and watermarks.

The 90 and 65 nm generation devices have been successfully integrated with copper and low permittivity materials with bulk dielectric constants in the range of 2.9–3.1. Numerous challenges have been overcome in achieving good device yields and reliability with this first generation of porous films. Further scaling of interconnects with forthcoming device generations will require the integration of even lower permittivity materials ($2.2 < k < 2.9$) that enable sizeable reductions in the interconnect capacitance. The inevitable degradation of mechanical properties with lower permittivity insulators poses formidable challenges to the integration of these materials into multilevel interconnect structures. Modifications to existing dual damascene fabrication schemes and processes may offer means to overcome some of the difficulties in the pattern transfer module. While the hierarchical reverse scaling approach of interconnect architecture offers some relief in the RC delay of the intermediate and global metal levels, continued scaling of physical dimensions demands the extension of lower permittivity dielectrics into the higher levels of multilevel interconnects. Lower permittivity materials also entail higher residual tensile stresses, increasing the risk of crack initiation and propagation. Engineering the packaging related processes [74–79] to accommodate devices with multiple levels of porous materials with low mechanical strength is a key to realize the successful implementation of low permittivity materials for intermediate and global interconnections.

The increasing resistivity of the copper with thinner and narrower lines raises serious concern about the extendibility of copper metallization. Efforts to reduce the relatively high volume of high resistivity barrier layers within the metal lines are hampered by the difficulties in reliable integration of atomic layer deposited barrier films with the copper metallization. Novel design approaches to comprehend and accommodate resistivity-dependent interconnect line width may be necessary to overcome the projected increase of line resistance as well as the increased variances in line resistance due to the pattern-dependent limitations of chemical–mechanical planarization processes. Ultimately, the extendibility of copper metallization with low permittivity insulators will depend on the economic and practical factors that dictate the new cost sensitive era of consumer devices.

2.3 Copper/Low-*k* Reliability

Copper metallization offers significant reliability improvement as compared with aluminum metallization, but also presents several new integration and reliability challenges. Since copper readily diffuses into silicon and most dielectrics, copper leads must be encapsulated with metallic (such as Ta and TaN) and dielectric (such as SiN and SiC) diffusion barriers to prevent electrical leakage between adjacent metal leads and degradation of transistor performance. Copper diffusion is greatly enhanced by electric fields imposed between adjacent leads during device operation ($\sim$1E5 V/cm), so absolute barrier integrity is crucial to long-term device reliability (see Figure 2.17). Validation of barrier reliability required a new test procedure—bias temperature stress (BTS)—now prevalently used by the industry [80–83].

The electromigration behavior of copper also differs from aluminum in that surface diffusion tends to dominate over grain boundary diffusion. This difference may be one reason that preliminary data show deterioration in copper reliability at very small feature sizes (i.e., a large ratio of surface area to cross-sectional area) [84]. From a resistivity standpoint, it is still important to maximize copper grain size to reduce grain boundary scattering, but as compared with aluminum, grain size plays a lesser role in determining copper's electromigration behavior.

Yet another key difference is that copper, unlike aluminum, does not form a self-limiting passivation oxide that provides protection from chemical attack. Indeed, oxidized copper generally has poor adhesion to metal and dielectric diffusion barriers, which leads to severe degradation in electromigration performance (see Figure 2.18). Thus, it is essential to chemically remove residual copper oxide before in situ deposition of a hermetic dielectric diffusion barrier (see pretreatment effects in Figure 2.19). As shown in Figure 2.20, dielectric barrier hermeticity is also important from the standpoint of preventing moisture absorption in low-*k* dielectrics, which can degrade both dielectric constant and electrical breakdown.

In addition to electromigration effects, inherent stress gradients in copper metallization can also lead to void migration even without the assistance of an electric field. Stress gradients in copper are a consequence of damascene processing (i.e., encapsulation) and subsequent thermal cycling (see Figure 2.21), with higher stress levels generally promoting the likelihood of void formation. Pre-existing

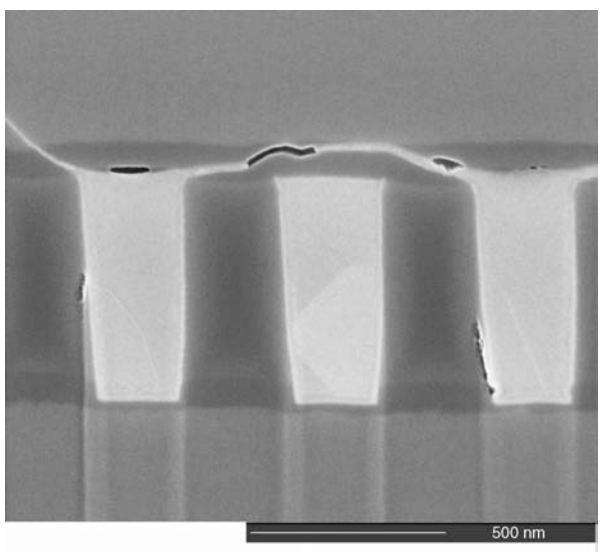


FIGURE 2.17 Example of dielectric barrier failure leading to copper diffusion and subsequent shorting between interconnects.

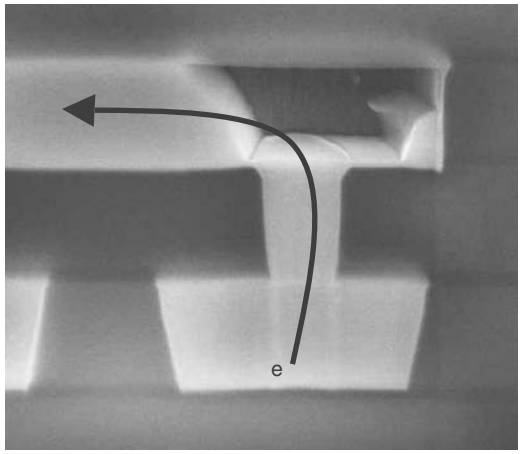


FIGURE 2.18 Example of void formed by copper electromigration.

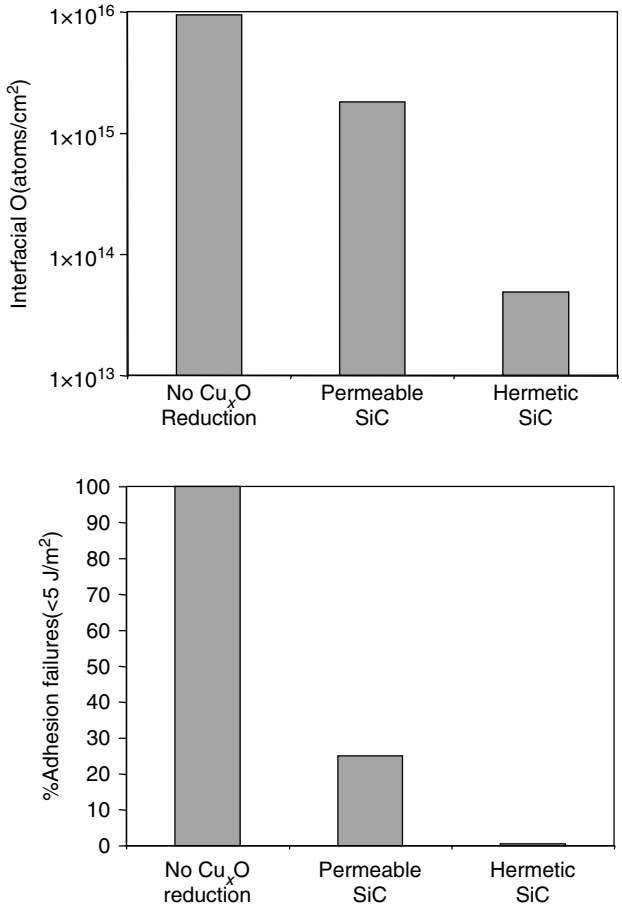


FIGURE 2.19 Improvements in adhesion and oxidation afforded by a hermetic dielectric barrier.

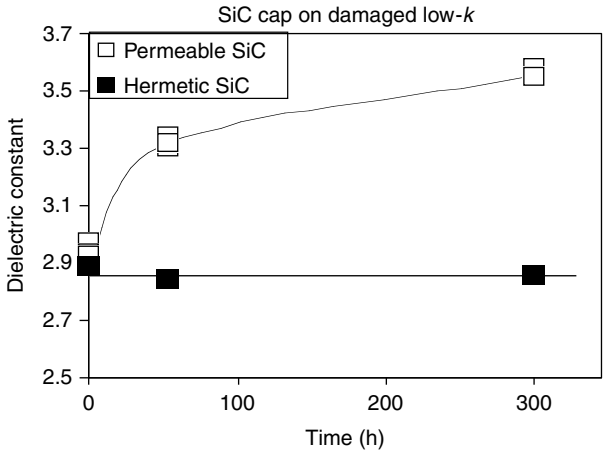


FIGURE 2.20 Improvement in low-*k* stability through the use of a hermetic dielectric barrier to prevent moisture absorption.

voids or interfaces with weak adhesion provide nucleation centers for void formation as shown in Figure 2.22. Thus, it is important to optimize dielectric etch/clean, chemical–mechanical planarization/post-CMP clean, copper/barrier metal deposition processes, and dielectric barrier deposition to produce chemically and mechanically stable interfaces and eliminate void nucleation centers. Stress management through judicious choice of copper plating chemistry and subsequent thermal anneal also plays a key role in controlling stress migration.

As metal line width and intrinsic barrier thickness decrease with scaling, copper containment becomes increasingly more problematic. Surface roughness, such as may be encountered on etched porous low-*k* dielectrics, brings additional concerns. While advanced PVD barriers have provided the conformality needed thus far, ALD barriers will likely be needed beyond the 32 nm technology node. Due to the domination of surface diffusion for copper electromigration, the barrier-to-copper interface quality will be a key factor in determining the success of any new approach.

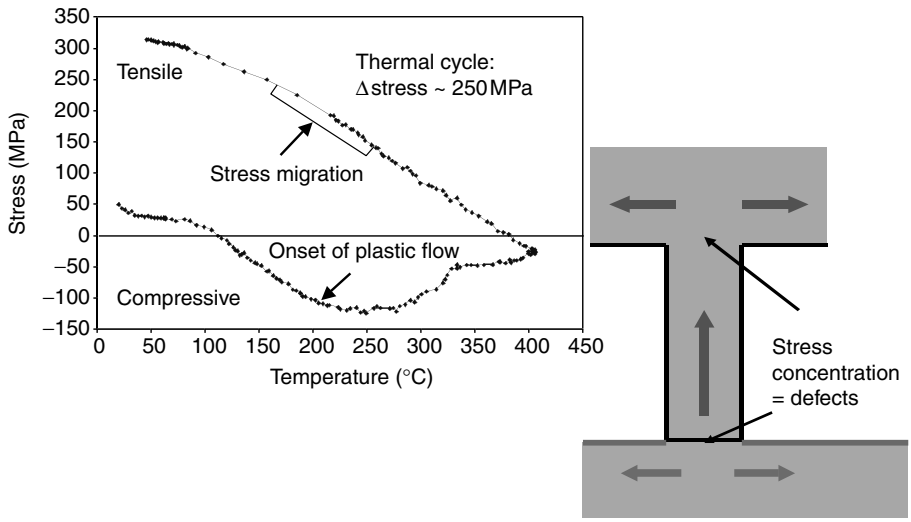


FIGURE 2.21 Copper stress transition as a function of thermal cycling.

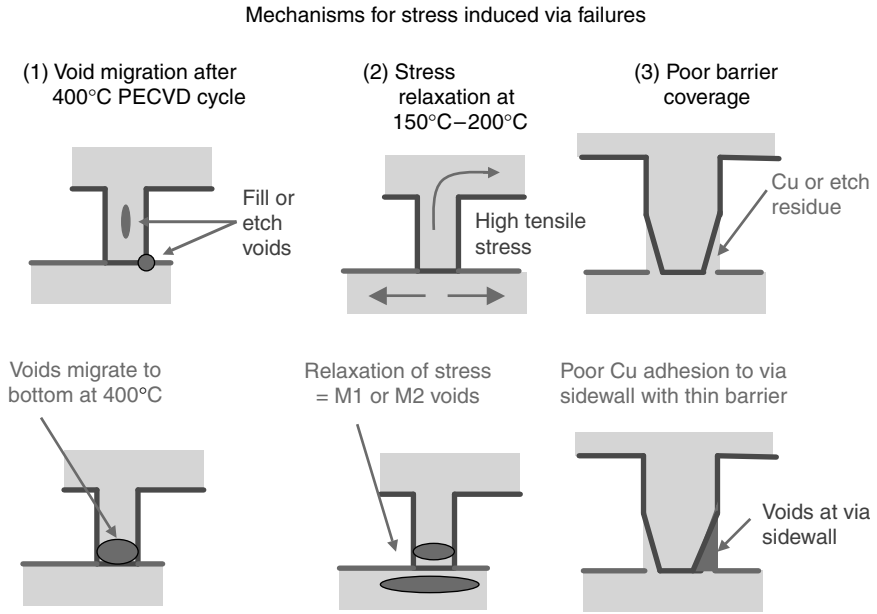

Mechanisms for stress induced via failures
(1) Void migration after 400°C PECVD cycle
(2) Stress relaxation at 150°C–200°C
(3) Poor barrier coverage
Fill or etch voids
High tensile stress
Cu or etch residue
Voids migrate to bottom at 400°C
Relaxation of stress = M1 or M2 voids
Poor Cu adhesion to via sidewall with thin barrier
Voids at via sidewall

FIGURE 2.22 Mechanisms for stress induced migration leading to via failures.

The integration of new low-*k* dielectrics needed for performance enhancement also bring numerous reliability issues that include thermally- or mechanically-induced cracking or adhesion loss, poor mechanical strength, moisture absorption, time-dependent behavior, lower dielectric breakdown voltage/TDDB, texture effects and poor thermal conductivity [85–96]. The reduced mechanical strength of porous low-*k* dielectrics is of particular concern for CMP and packaging; new planarization and packaging approaches will be needed to accommodate these mechanically weaker materials.

2.4 Conclusion

The integration of copper metallization with low-*k* dielectrics has undoubtedly served as an enabler for continued device scaling. By leveraging copper's lower resistivity as compared with aluminum, both latency and power distribution in integrated circuits have been improved, and the enhanced reliability of copper interconnects has also supported the use of higher current densities in scaled devices. Further improvements in latency as well as crosstalk have been afforded by the integration of low-*k* dielectrics with copper metallization, but the need for even lower-*k* dielectrics beyond the 45 nm technology node presents significant materials and process integration challenges. While the copper damascene process has enabled higher interconnect density and improved yield, substantial improvements in barrier/seed deposition, copper electroplating, and CMP planarization will be necessary for continued scaling; new materials and processes, such as ALD barriers, will likely be required at the 22 nm technology node and beyond. As metal line widths continue to scale downward, the additional electrical limitations of electron scattering come into play, and new materials and/or processes, which minimize scattering must be implemented to stem the tide of increasing copper "effective" resistivity. A further consequence of smaller metal line widths is a heightened sensitivity to stress migration, and new materials and process architectures will likely be required to maintain robust reliability. While significant process integration challenges lie ahead, the successful implementation of copper and low-*k* dielectrics in multiple generations of volume production provides both the confidence and learning to continue moving forward as an essential part of the IC scaling engine.

References

1. Edelstein, D., J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, P. Roper, et al. "Full Copper Wiring in a Sub-0.25 μm CMOS ULSI Technology." *Tech. Dig. IEEE Int. Electron Devices Meeting* (1997): 773–6.
2. Venkatesan, S., A. V. Gelatos, V. Misra, B. Smithe, R. Islam, J. Cope, B. Wilson, et al. "A High Performance 1.8 V, 0.20 μm CMOS Technology with Copper Metallization." *Tech. Dig. IEEE Int. Electron Devices Meeting* (1997): 769–72.
3. Edelstein, D. C., G. A. Sai-Halasz, and Y.-J. Mii. "VLSI On-Chip Interconnection Performance Simulations and Measurements." *IBM J. Res. Dev.* 39, no. 4 (1995): 383.
4. Stamper, A. K., T. L. McDevitt, and S. L. Luce. "Sub-0.25-micron Interconnection Scaling: Damascene Copper Versus Subtractive Aluminum." *9th IEEE/SEMI Adv. Semicond. Manuf. Conf. Workshop* (1998): 337–46.
5. Deutsch, A., H. Harrer, C. W. Surovic, G. Hellner, D. C. Edelstein, R. D. Goldblatt, G. A. Biery, et al. "Functional High-Speed Characterization and Modeling of a Six-Layer Copper Wiring Structure and Performance Comparison with Aluminum On-Chip Interconnections." *Tech. Dig. IEEE Int. Electron Devices Meeting* (1998): 295–8.
6. Smith, B., S. Blackley, R. Carter, S. Chheda, P. Crabtree, D. Farber, M. Gall, et al. "A Comparison of via Overetch Variations Between Conventional Al-W and Dual-Inlaid Copper Integrations." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (1999): 106–8.
7. Rosenberg, R., D. C. Edelstein, C.-K. Hu, and K. P. Rodbell. "Copper Metallization for High Performance Silicon Technology." *Annu. Rev. Mater. Sci.* 30 (2000): 229–62.
8. Atwood, G. and S. Lai. "Future Directions and Challenges for Flash Memory Scaling." In *Proceedings of the International Reliability Physics Symposium*, Tutorial Notes, 2004.
9. MacGillivray G. "90 nm StrataFlash Eyes Mobile." *EE Times* Nov (2005).
10. Tao, J., N. W. Cheung, and C. Hu. "Electromigration Characteristics of Copper Interconnects." *IEEE Electron Device Lett.* 14, no. 5 (1993): 249–51.
11. Ogawa, E., K. Lee, V. Blaschke, and P. Ho. "Electromigration Reliability Issues in Dual-Damascene Cu Interconnects." *IEEE Trans. Reliability* 51, no. 4 (2002): 403–19.
12. Zielinski, E., S. Russell, R. List, A. Wilson, C. Jin, K. Newton, J. Lu, et al. "Damascene Integration of Copper and Ultra-Low-*k* Xerogel for High Performance Interconnects." *Tech. Dig. IEEE Int. Electron Devices Meeting* (1997): 936–8.
13. Crowder, S., S. Greco, H. Ng, E. Barth, K. Beyer, G. Biery, J. Connolly, et al. "A 0.18 μm High-Performance Logic Technology." *Symp. VLSI Tech. Dig. Tech. Papers* (1999): 105–6.
14. Goldblatt, R. D., B. Agarwala, M. B. Anand, E. P. Barth, G. A. Biery, Z. G. Chen, S. Cohen, et al. "A High Performance 0.13 μm Copper BEOL Technology with Low-*k* Dielectric." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2000): 261–3.
15. Takao, Y., H. Kudo, J. Mitani, Y. Kotani, S. Yamaguchi, K. Yoshie, M. Kawano, et al. "A 0.11 μm CMOS Technology with Copper and Very-Low-*k* Interconnects for High-Performance System-on-a-Chip Cores." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2000): 559–62.
16. Young, K. K., S. Y. Wu, C. C. Wu, C. H. Wang, C. T. Lin, J. Y. Cheng, M. Chiang, et al. "A 0.13 μm CMOS Technology with 193 nm Lithography and Cu/Low-*k* for High Performance Applications." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2000): 563–6.
17. Tyagi, S., M. Alavi, R. Bigwood, T. Bramblett, J. Brandenburg, W. Chen, B. Crew, et al. "A 130 nm Generation Logic Technology Featuring 70 nm Transistors, Dual V_T Transistors and. 6 layers of interconnects." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2000): 567–70.
18. Perera, A., B. Smith, N. Cave, M. Sureddin, S. Chheda, R. Singh, R. Islam, et al. "A Versatile 0.13 μm CMOS Platform Technology Supporting High Performance and Low Power Applications." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2000): 571–4.
19. Jeng, S. P., M.-C. Chang, and R. H. Havemann. "Process Integration and Manufacturing Issues for High Performance Interconnect." *MRS Symp. Proc. Adv. Metal. Devices Circuits* (1994): 25–31.
20. Bohr, M. "Interconnect Scaling—The Real Limiter to High Performance ULSI." *Tech. Dig. IEEE Int. Electron Devices Meeting* (1995): 241–4.

21. Rahmat, K., O. S. Nakagawa, S.-Y. Oh, J. Moll, and W. T. Lynch. "A Scaling Scheme for Interconnect in Deep-Submicron Processes." *Tech. Dig. IEEE Int. Electron Devices Meeting* (1995): 245–8.
22. Yamashita, K. and S. Odanaka. "Impact of Crosstalk on Delay Time and a Hierarchy of Interconnects." *Tech. Dig. IEEE Int. Electron Devices Meeting* (1998): 291–4.
23. Takahashi, S., M. Edahiro, and Y. Hayashi. "Interconnect Design Strategy: Structures, Repeaters and Materials Toward 0.1 μm ULSIs with a Giga-Hertz Clock Operation." *Tech. Dig. IEEE Int. Electron Devices Meeting* (1998): 833–6.
24. Havemann, R. H. and J. A. Hutchby. "High-Performance Interconnects: An Integration Overview." *Proc. IEEE* 89, no. 5 (2001): 586–601.
25. Fisher, P. and R. Nesbitt. "The Test of Time Clock-Cycle Estimation and Test Challenges for Future Microprocessors." *IEEE Circuits Devices* (1998): 37–44.
26. "Interconnect." In *International Technology Roadmap for Semiconductors*, published by the Semiconductor Industry Association, 2005.
27. "Executive Summary." *International Technology Roadmap for Semiconductors*, published by the Semiconductor Industry Association, 2005.
28. Shih, W.-Y., M.-C. Chang, R. H. Havemann, and J. Levine. "Implications and Solutions for Joule Heating in High Performance Interconnects Incorporating Low-*k* Dielectrics." *Symp. VLSI Tech. Dig. Tech. Papers* (1997): 83–4.
29. Floyd, B. A. and K. K. O. "The Projected Power Consumption of a Wireless Clock Distribution System and Comparison to Conventional Systems." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (1999): 248–51.
30. Daly, W. J. "Interconnect-Limited VLSI architecture." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (1999): 15–7.
31. Venkatesan, R., J. Davis, K. Bowman, and J. Meindl. "Optimal Repeater Insertion for *n*-Tier Multilevel Interconnect." *Tech. Dig. IEEE Int. Electron Devices Meeting* (2000): 132–4.
32. Kapur, P., J. P. McVittie, and K. C. Saraswat. "Realistic Copper Interconnect Performance with Technological Constraints." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2001): 233–5.
33. Davis, J. A., R. Venkatesan, A. Kaloyeros, M. Beylansky, S. J. Souri, K. Banerjee, K. C. Saraswat, A. Rahman, R. Reif, and J. D. Meindl. "Interconnect Limits on Gigascale Integration (GSI) in the 21st Century." *Proc. IEEE* 89, no. 3 (2001): 305–24.
34. Banerjee, K. and A. Mehrotra. "Power Dissipation Issues in Interconnect Performance Optimization for Sub-180 nm Designs." *Tech. Dig. Symp. VLSI Circuits* (2002): 12–5.
35. Deodhar, V. V. and J. A. Davis. "Voltage Scaling and Repeater Insertion for High-Throughput Low-Power Interconnects." *Tech. Dig. Int. Symp. Circuits Syst.* (2003): V-349–V-352.
36. Dasgupta, P. "Revisiting VLSI Interconnects in Deep Sub-Micron: Some Open Questions." *Tech. Dig. Int. Conf. VLSI Des.* (2005): 615–20.
37. Nagaraj, N. S., W. R. Hunter, R. Chidambaram, T. Y. Garibay, U. Narasimha, A. Hill, and H. Shichijo. "Impact of Interconnect Technology Scaling on SOC Design Methodologies." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2005): 71–3.
38. Wang, K. N., J. M. Adam, and P. A. Dziekowicz. "Electrical Performance Trade-Offs in Ball Grid Array Package Designs." *Tech. Dig. Seventeenth IEEE/CPMT Int. Electron. Manuf. Technol. Symp.* (1995): 416.
39. Huang, C. C., D. Secker, L. Yang, J. Feng, and N. Jain. "Design and Characterization of a High-Performance Wire-Bond Ball-Grid-Array Package." *Tech. Dig. IEEE/SEMI Int. Electron. Manuf. Technol. Symp.* (2002): 245–9.
40. Ma, Y. Y., D. Y. R. Chong, C. K. Wang, and A. Y. S. Sun. "Development of Ball Grid Array Packages with Improved Thermal Performance." *Tech. Dig. Electron. Packaging Technol. Conf.* 2 (2005): 6.
41. Kajita, A., T. Usui, M. Yamada, E. Ogawa, T. Katata, A. Sakata, H. Miyajima, et al. "Highly Reliable Cu/Low-*k* Dual-Damascene Interconnect Technology with Hybrid (PAE/SiOC) Dielectrics for 65 nm-Node Performance eDRAM." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2003): 9–11.
42. Edelstein, D., C. Davis, L. Clevenger, M. Yoon, A. Cowley, T. Nogami, H. Rathore, et al. "Reliability Yield, and Performance of a 90 nm SOI/Cu/SiCOH Technology." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2004): 214–6.

43. Jan, C. H., N. Anand, C. Allen, J. Bielefeld, M. Buehler, V. Chikamane, K. Fischer, et al. "A 90 nm High Volume Manufacturing Logic Technology Featuring Cu Metallization and CDO Low-*k* ILD Interconnects on 300 mm Wafers." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2004): 205–7.
44. Jeng, C. C., W. K. Wan, H. H. Lin, K. H. Tang, I. C. Kao, H. C. Lo, K. S. Chi, et al. "BEOL Process Integration of 65 nm Cu/Low-*k* Interconnects." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2004): 199–201.
45. Fukasawa, M., S. Lane, M. Angyal, K. Chanda, F. Chen, C. Christiansen, J. Fitzsimmons, et al. "BEOL Process Integration with Cu/SiCOH ($k=2.8$) Low-*k* Interconnects at 65 nm Groundrules." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2005): 9–11.
46. Chatterjee, A., J. Yoon, S. Zhao, S. Tang, K. Sadra, S. Crank, H. Mogul, et al. "A 65 nm CMOS Technology for Mobile and Digital Signal Processing Applications." *Tech. Dig. IEEE Int. Electron Devices Meeting* (2004): 665–8.
47. Fox, R., O. Hinsinger, E. Richard, E. Sabouret, T. Berger, C. Goldberg, A. Humbertw, et al. "High Performance $k=2.5$ ULK Backend Solution Using an Improved TFHM Architecture, Extendible to the 45 nm Technology Node." *Tech. Dig. IEEE Int. Electron Devices Meeting* (2005): 81–4.
48. Matsunaga, N., N. Nakamura, K. Higashi, H. Yamaguchi, T. Watanabe, K. Akiyama, S. Nakao, et al. "BEOL Process Integration Technology for 45 nm Node Porous Low-*k*/Copper Interconnects." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2005): 6–8.
49. Soo, C. P., S. Valiyaveettil, A. Huan, A. Wee, C. A. Ting, M. H. Fan, A. J. Bourdillon, and H. C. Lap. "Enhancement or Reduction of Catalytic Dissolution Reaction in Chemically Amplified Resists by Substrate Contaminants." *IEEE Trans. Semicond. Manuf.* 12, no. 4 (1999): 462–9.
50. Dixit, G. "Low-*k*/Cu Integration." In *Proceedings of the International Reliability Physics Symposium*, Tutorial Notes, 2004.
51. Nogami, T., S. Lane, M. Fukasawa, K. Ida, M. Angyal, K. Chanda, F. Chen., et al. "Low-*k*/copper Integration Scheme Suitable for ULSI Manufacturing from 90 to 45 nm nodes." *Proc. SPIE* 6002 (2005): 90–104.
52. Furusawa, T., S. Machida, D. Ryuzaki, K. Sameshima, T. Ishida, K. Ishikawa, N. Miura, N. Konishi, T. Saito, and H. Yamaguchi. "Dual-Damascene Cu/Low-*k* Interconnect Fabrication Scheme Using Dissoluble Hard Mask Material." *J. Electrochem. Soc.* 153, no. 2 (2006): G160–3.
53. Hinsinger, O., R. Fox, E. Sabouret, C. Goldberg, C. Verove, W. Besling, P. Brun, et al. "Demonstration of an Extendable and Industrial 300 mm BEOL Integration for the 65-nm Technology Node." *Tech. Dig. IEEE Int. Electron Devices Meeting* (2004): 317–20.
54. Liang, M. S. "Challenges in Cu/Low-*k* Integration." *Tech. Dig. IEEE Int. Electron Devices Meeting* (2004): 313–6.
55. Miyajima, H., K. Watanabe, K. Fujita, S. Ito, K. Tabuchi, T. Shimayama, K. Akiyama, et al. "Challenge of Low-*k* Materials for 130, 90, 65 nm Node Interconnect Technology and Beyond." *Tech. Dig. IEEE Int. Electron Devices Meeting* (2004): 329–32.
56. Dalton, T. J., N. Fuller, C. Tweedie, D. Dunn, C. Labelle, S. Gates, M. Colburn, et al. "Ash-Induced Modification of Porous and Dense SiCOH Inter-Level-Dielectric (ILD) Materials During Damascene Plasma Processing." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2004): 154–6.
57. Posseme, N., C. Maurice, Ph.Brun, E. Ollier, M. Guillermet, C. Verove, T. Berger, R. Fox, and O. Hinsinger. "New Etch Challenges for the 65-nm Technology Node Low-*k* Integration Using an Enhanced Trench First Hard Mask Architecture." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2006): 36–8.
58. Iacopi, F., M. Stucchi, O. Richard, and K. Maex. "Electrical Equivalent Sidewall Damage in Patterned Low-*k* Dielectrics." *Electrochem. Solid-State Lett.* 7, no. 4 (2004): G79–82.
59. Struyf, H., D. Hendrickx, J. Van-Olmen, F. Iacopi, O. Richard, Y. Travaly, M. Van-Hove, W. Boullart, and S. Vanhaelemeersch. "Low-Damage Damascene Patterning of SiOC(H) Low-*k* Dielectrics." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2005): 30–2.
60. Baklanov, M. R., Q. T. Le, E. Kesters, F. Iacopi, J. VanAelst, H. Struyf, W. Boullart, S. Van haelemeersch, and K. Maex. "Challenges of Clean/Strip Processing for Cu/Low-*k* Technology." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2004): 187–9.

61. Clevenger, L., M. Yoon, D. Edelstein, A. Cowley, C. Davis, B. Agarwala, P. Biolsi, et al. "90 nm SiCOH Technology in 300 mm Manufacturing." *Pro. Adv. Metal. Conf.* (2004): 27–36.
62. Alers, G. B., R. T. Rozbicki, G. J. Harm, S. K. Kailasam, G. W. Ray, and M. Danek. "Barrier-First Integration for Improved Reliability in Copper Dual Damascene Interconnects." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2003): 27–9.
63. Edelstein, D., C. Uzoh, C. Cabral, P. DeHaven, P. Buchwalter, A. Simon, E. Cooney, et al. "A High Performance Liner for Copper Damascene Interconnects." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2001): 9–11.
64. Chenand, F., and D. Gardner. "Influence of Line Dimensions on the Resistance of Cu Interconnections." *IEEE Electron Device Lett.* 19 (1998): 508–10.
65. Rossnagel, S. M., and S. T. Kuan. "Alteration of Cu Conductivity in the Size Effect Regime." *J. Vac. Sci. Technol.* B22 (2004): 240–7.
66. Zantye, P. B., A. Kumar, and A. K. Sikder. "Chemical Mechanical Planarization for Microelectronics Applications." *Mater. Sci. Eng. R.* 45 (2004): 89–220.
67. Babu, S. V., Y. Li, and A. Jindal. "Chemical–Mechanical Planarization of Cu and Ta." *JOM* 53 (2001): 50–2.
68. Leduc, P., M. Savoye, S. Maitrejean, D. Scevola, V. Jousseaume, and G. Passemard. "Understanding CMP-Induced Delamination in Ultra Low-*k*/Cu Integration." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2005): 209–11.
69. Kondo, S., B. U. Yoon, S. G. Lee, S. Tokitoh, K. Misawa, T. Yoshie, N. Ohashi, and N. Kobayashi. "Damage-Free CMP Towards 32 nm-Node Porous Low-*k* ($k=1.6$)/Cu Integration." *Tech. Dig. VLSI Technol. Symp.* (2004): 68–9.
70. Pallinti, J., S. Lakshminarayanan, W. Barth, P. Wright, M. Lu, S. Reder, L. Kwak, W. Catabay, D. Wang, and F. Ho. "An Overview of Stress Free Polishing of Cu with Ultra Low-*k* ($k<2.0$) Films." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2003): 83–5.
71. Kondo, S., S. Tokitoh, B. U. Yoon, A. Namiki, N. Ohashi, K. Misawa, S. Sone, et al. "Low-Pressure CMP for Reliable Porous Low-*k*/Cu Integration." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2003): 86–8.
72. Guyer, E. P., and R. H. Dauskardt. "Effect of CMP Slurry Environments on Subcritical Crack Growth in Ultra Low-*k* Dielectric Materials." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2003): 89–91.
73. Kondo, S., B. U. Yoon, S. Tokitoh, K. Misawa, S. Sone, H. J. Shin, N. Ohashi, and N. Kobayashi. "Low-Pressure CMP for 300-mm Ultra Low-*k* ($k=1.6$–1.8)/Cu Integration." *Tech. Dig. IEEE Int. Electron Devices Meeting* (2003): 641–4.
74. Mercado, L. L., C. Goldberg, S. M. Kuo, T. Y. T. Lee, and S. Pozder. "Analysis of Flip-Chip Packaging Challenges on Copper Low-*k* Interconnects." *Proc. 53rd Electron. Components Technol. Conf.* (2003): 1784–90.
75. Wang, G., C. Merrill, J. H. Zhao, S. K. Groothuis, and P. S. Ho. "Packaging Effects on Reliability of Cu/Low-*k* Interconnects." *IEEE Trans. Device Mater. Reliability* 2 (2003): 119–28.
76. Mercado, Lei L., S. M. Kuo, C. Goldberg, and D. Frear. "Impact of Flip-Chip Packaging on Copper/Low-*k* Structures." *IEEE Trans. Adv. Packaging* 26 (2003): 433–40.
77. Goldberg, C., S. Downey, V. Fiori, R. Fox, K. Hess, O. Hinsinger, A. Humbert, et al. "Integration of a Mechanically Reliable 65-nm Node Technology for Low-*k* and ULK Interconnects with Various Substrate and Package Types." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2005): 3–5.
78. Tagami, M., H. Ohtake, M. Abe, F. Ito, T. Takeuchi, K. Ohto, and T. Usami. "Comprehensive Process Design for Low-Cost Chip Packaging with Circuit-Under-Pad (CUP) Structure in Porous SiCOH Film." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2005): 12–4.
79. Uchibori, C., X. Zhang, P. S. Ho, and T. Nakamura. "Effects of Chip-Package Interaction on Mechanical Reliability of Cu Interconnects for 65 nm Technology Node and Beyond." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2006): 196–8.
80. Raghavan, G., C. Chiang, P. B. Anders, S. Tzeng, R. Villasol, G. Bai, T. Bohr, and D. B. Fraser. "Diffusion of Copper Through Dielectric Films Under Bias Temperature Stress." *Thin Solid Films* 262 (1995): 168–76.

81. Tsu, R., J. W. McPherson, and W. R. McKee. "Leakage and Breakdown Reliability Issues Associated with Low-*k* Dielectrics in a Dual-Damascene Cu Process." *Proc. IEEE Int. Reliability Phys. Symp.* (2000): 348–53.
82. Noguchi, J., N. Miura, M. Kubo, T. Tsuyoshi, H. Yamaguchi, N. Hamada, K. Makabe, R. Tsuneda, and K. Takeda. "Cu-Ion-Migration Phenomena and its Influence on TDDB Lifetime in Cu Metallization." *Proc. IEEE Int. Reliability Phys. Symp.* (2003): 287–92.
83. Ogawa, E. T., J. Kim, G. S. Haase, H. C. Mogul, and J. W. McPherson. "Leakage, Breakdown and TDDB Characteristics of Porous Low-*k* Silica Based Interconnect Dielectrics." *Proc. IEEE Int. Reliability Phys. Symp.* (2003): 166–72.
84. Hu, C.-K., R. Rosenberg, H. S. Rathore, D. B. Nguyen, and B. Agarwala. "Scaling Effect on Electromigration in On-Chip Cu Wiring." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (1999): 267–9.
85. Hu, C.-K., R. Rosenberg, and K. L. Lee. "Electromigration Path in Cu Thin-film Lines." *Appl. Phys. Lett.* 74 (1999): 2945.
86. Kuan, T. S., C. K. Inoki, G. S. Oehrlein, K. Rose, Y.-P. Zhao, G. C. Wang, S. M. Rossnagel, and C. Cabral. "Fabrication and Performance Limits of Sub-0.1 Micrometer Cu Interconnects." *Mater. Res. Soc. Symp. Proc.* 612 (2000): D7.1.1.
87. Hayashi, M., S. Nakano, and T. Wada. "Dependence of Copper Interconnect Electromigration Phenomenon on Barrier Metal Materials." *Microelectron. Reliability* 43 (2003): 1545.
88. Li, B., T. D. Sullivan, T. C. Lee, and D. Badami. "Reliability Challenges for Copper Interconnects." *Microelectron. Reliability* 44 (2004): 365–80.
89. Liniger, E. G., C.-K. Hu, L. M. Gignac, and A. Simon. "Effect of Liner Thickness on Electromigration Lifetime." *J. Appl. Phys.* 93, no. 12 (2003): 9576–82.
90. Li, B., T. Sullivan, and T. C. Lee. "Line Depletion Electromigration Characterization of Cu Interconnects." *IEEE Trans. Device Mater. Reliability* 4, no. 1 (2004): 80–5.
91. Fischer, A. H., A. Glasow, S. Penka, and F. Ungar. "Electromigration Failure Mechanism Studies on Copper Interconnects." *Tech. Dig. IEEE Int. Interconnect Tech. Conf.* (2002): 139–41.
92. Lane, M. W., E. G. Liniger, and J. R. Lloyd. "Relationship Between Interfacial Adhesion and Electromigration in Cu Metallization." *J. Appl. Phys.* 93 (2003): 1417.
93. Kimura, M. "Oxide Breakdown Mechanism and Quantum Physical Chemistry for Time-Dependent Dielectric Breakdown." *Proc. IEEE Int. Reliability Phys. Symp.* (1997): 190–200.
94. Alers, G. B., M. Sanganeria, R. Shaviv, G. Kooi, K. Jow, and G. W. Ray. "Failure Mechanisms in Dielectric Barriers." *Proc. Adv. Metal. Conf.* (2003).
95. Noguchi, J., N. Ohashi, T. Jimbo, H. Yamaguchi, K. Takeda, and K. Hinode. "Effect of NH-Plasma Treatment and CMP Modification TDDB Improvement in Cu Metallization." *IEEE Trans. Electron Devices* 48, no. 7 (2001): 1340–5.
96. Wu, W., X. Duan, and J. S. Yuan. "A Physics Model of Time-Dependent Dielectric Breakdown in Cu Metallization." *Proc. IEEE Int. Reliability Phys. Symp.* (2003): 773–6.

3

Silicon Materials

3.1 Introduction **3**-1

3.2 Silicon Crystal Growth Processes **3**-2
Float Zone Silicon Growth • Czochralski Silicon Growth

3.3 Characteristics of Czochralski Silicon Growth **3**-5
Thermal Characteristics for Dislocation-Free Growth • Impurity Incorporation • Grown-In Microdefects

3.4 Trends in Large Diameter Silicon Growth **3**-38
Evolution in Crystal Diameter • Continuous Czochralski Silicon Growth

3.5 Wafer Preparation **3**-45
Slicing • Chemical Etching • Edge Rounding • Lapping/Grinding • Polishing • Cleaning

3.6 Epitaxial Growth **3**-49
Silicon Epitaxial Wafer • Heteroepitaxy • Selective Epitaxial Growth • Strained Silicon Epitaxy

3.7 Oxygen Behavior in Silicon Processing **3**-61
Oxygen Precipitation Kinetics • Oxygen Precipitation in p^+ and n^+ • Effects of Oxygen Precipitation on Device Processing • Denuded Zone and Oxygen Precipitation by Controlled Vacancy Concentration

3.8 Other and New Applications of Silicon Materials **3**-68
Nitrogen Doping and Its Effects on CZ Silicon • High Resistivity Silicon

3.9 Summary **3**-70

References **3**-71

Wen Lin
Consultant

Howard Huff
SEMATECH

3.1 Introduction

Silicon is the major semiconductor material used in solid state electronics. Silicon in the form of a single crystal wafer is the basic building block for the integrated circuit (IC) fabrication. To keep pace with the growth in IC processing technology, chip size and circuit complexity, silicon crystal and wafers have to be prepared with continued increases in diameter and improvements in perfection. Modern ultra large scale integration (ULSI) ICs, fabricated with design rules approaching 60 nm, have to depend on the availability of highly perfect single crystals, which are prepared exclusively from silicon pulled from the melt by the Czochralski (CZ) technique. The CZ silicon material prepared today must meet the challenges imposed by the stringent requirements of the starting materials for current and future design rule generations, as guided by the Starting Materials Road Map of the International Technology Roadmap for Semiconductor (ITRS).

The purpose of this chapter is to review the fundamental aspects of silicon technology and to discuss the advances in understanding of the science and engineering of this technology with respect to the challenges in the material requirements for ULSI fabrication.

3.2 Silicon Crystal Growth Processes

3.2.1 Float Zone Silicon Growth

The float zone method (FZ) is based on the zone-melting principle [1] and was invented by Theuerer [2]. Figure 3.1 shows a schematic of the FZ process. A polysilicon rod is mounted vertically inside a growth chamber under vacuum or an inert atmosphere. A needle-eye coil provides radio frequency (RF) power to the rod causing it to melt and maintain a narrow, stable molten zone by balancing the surface tension and gravitational forces. The levitation effect of the RF field helps to support a large molten zone. As the molten zone is moved along upward the polysilicon rod, the molten silicon solidifies into a single crystal and, simultaneously, the material is purified. To initiate the growth, in the bottom-seed FZ, the seed crystal ($\sim$10 mm in diameter) is brought up from below to make contact with the drop of melt formed at the tip of the poly rod. A necking process is carried out to establish dislocation-free feature before the "neck" is allowed to increase in diameter, to form a taper and to reach the desired diameter for steady-state body growth. During this process, the shape of the molten zone and crystal diameter is monitored by infrared sensors and are adjusted by the RF power input to the coil and travel speed. Details of FZ technology are discussed by Keller and Muhlbauer [3]. Current FZ technology can produce a high quality FZ silicon up to 200 mm in diameter in production quantities.

Float zone crystals are doped by adding the doping gas phosphine (PH_3) or diborane (B_2H_6) to the inert gas for *n*- and *p*-type, respectively. Polysilicon rods for FZ growth may also be doped in the gas phase and dopant redistribution by zone melting. Since the doping is by gas phase interaction with the molten silicon, axial dopant uniformity is achieved. However, due to the very nature of FZ growth configuration, the small "hot zone" lacks thermal symmetry. As a result, temperature fluctuations, remelting phenomenon, and dopant segregation cause FZ silicon to display more microscopic dopant inhomogeneity or dopant striations than observed in CZ silicon. Severe dopant micro-inhomogeneity can be circumvented for *n*-type FZ via neutron transmutation doping (NTD) [4]. In NTD, a high purity FZ crystal is subjected to thermal neutron bombardment, causing some of the silicon isotope ^{30}Si ($\sim$3.1% of Si) to form the unstable isotope ^{31}Si, which decays to form the stable phosphorus isotope ^{31}P,

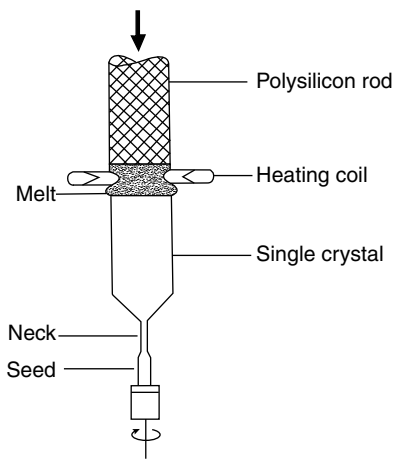


FIGURE 3.1 Schematic of a float zone (FZ) silicon growth arrangement.

such that

$$^{30}\mathrm{Si}(n,\gamma)\rightarrow{}^{31}\mathrm{Si}\rightarrow{}^{31}\mathrm{P}+\beta^{-} \quad (2.6\ \mathrm{hr}) \tag{3.1}$$

Since neutron bombardment (both thermal and fast neutrons) induces radiation damages, the irradiated crystal must be annealed at about 700°C for defect annihilation and to restore resistivity due to the phosphorus doping. In the FZ silicon with NTD, the dopant striations are greatly reduced. However, the NTD method is only feasible for high resistivity and phosphorus-doped FZ. Low resistivity doping of FZ by NTD would require excessively long irradiation (more lattice damages) and is not feasible. The NTD process for *p*-type doping is not available.

Unlike CZ growth, the silicon molten zone in FZ growth is not in contact with any substances except the ambient gas, which may contain doping gas. Therefore, the FZ silicon can easily achieve much higher purity and resistivity (resistivity ranges from a few tens to a few thousands ohm–cm) than the CZ silicon (generally, <50 ohm–cm). A large volume of FZ silicon is used in the fabrication of semiconductor power devices and infrared sensors. However, its application in the microelectronic IC fabrications is rather limited. The main reason is FZ's low interstitial oxygen content. The oxygen content of FZ crystal is in the range of 10^{16} atoms/cm^3 [5], which is two orders of magnitude below that of conventional CZ silicon and is far below the solid solubility limit at IC's thermal processing temperatures. Consequently, FZ silicon lacks the ability of oxygen precipitation and associated internal gettering potential. Furthermore, the "residual level" of oxygen in FZ also results in a lower mechanical strength than the CZ, in terms of its ability to withstand thermal stress and to suppress slip, wafer bow and warp during high-temperature processing [6,7]. The presence of "sufficient" interstitial oxygen in CZ silicon appears to have strain hardening effects in the silicon lattice and serves as obstacles for dislocation initiation/propagation. Numerous studies [8,9] have shown that the difference in mechanical strength can be attributed to the difference in the oxygen content and associated defects. These are the main reasons that CZ silicon is almost exclusively used for IC processing. In order to overcome FZ's shortcomings, oxygen [10] and nitrogen [11] doping during FZ growth have been reported. It was shown that doping concentrations of $1–1.5\times10^{17}$ atoms/cm^3 for oxygen or 1.5×10^{15} atoms/cm^3 for nitrogen can significantly increase FZ's strength. Nitrogen-doped FZ proves useful in microelectronics device fabrication, where oxygen-free material is desirable [12].

3.2.2 Czochralski Silicon Growth

Czochralski pulling method is named after Czochralski [13]. However, the pull-from-melt method widely employed today was developed by Teal and Little [14]. Small diameter CZ silicon crystals were first used in the early 1960s for IC fabrication (and earlier for discrete Ge and Si transistor fabrication). The development of the dislocation-free growth technique and automatic diameter control (ADC) in the late 1960s led to a rapid growth in silicon crystal diameter and charge/grower size during the last two decades. Today, 300 mm silicon crystals, weighing over 200 kg are common and the 450 mm era is already under development (initial research considered 400 mm as well as 450 mm-diameter) [109]. In ULSI era, the emphasis on the CZ silicon crystals for IC's is the uniformity of dopant, oxygen incorporation, and microdefect control. Significant research and development efforts are devoted to these aspects, in order to satisfy the stringent requirements of IC processing for deep sub-micron ICs. The CZ silicon growth process is controlled by many variables. It involves aspects of heat transfers and material phase transitions. Modern CZ silicon growth equipment transformed the "art" of crystal growth into an engineering discipline. Although most of the steps in the growth process can be monitored and controlled by a computer/microprocessor, many critical steps nevertheless still require the judgment and attention of an experienced operator. Therefore, significant differences can still exist in the properties of the crystals grown from different grower designs and growth processes.

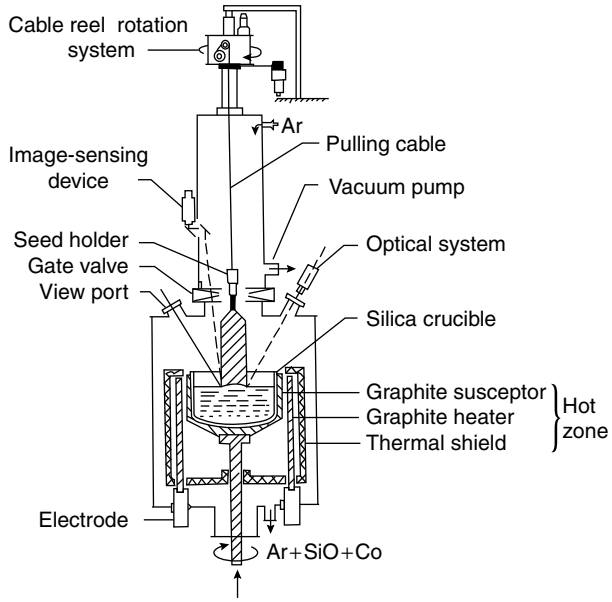


FIGURE 3.2 Schematic of a typical Czochralski silicon growing system. (From Abe, T., *VLSI Electronics Microstructure Science*, ed. Einspruch, N. G. and Huff, H. R., Academic Press, New York, Vol. 12, 1985. 3. Reproduced with permission from Elsevier.)

Figure 3.2 shows a schematic of a typical modern CZ grower for large diameter silicon growth. Major components include the so-called hot zone, crystal-pull/rotation and crucible-lift/rotation mechanisms, diameter, and temperature sensing devices. The puller is operated under a reduced pressure inert ambient (typically 18–20 torr). Czochralski silicon growth consists of three major processes: (1) seeding and necking, (2) body growth, and (3) tang growth/termination. The necking process is one of the most critical steps in dislocation-free CZ growth; the body growth cannot be initiated without a successful necking process. The necessity of the procedure and principles of necking are given below and are also common to the FZ growth discussed in Section 3.2.1.

The CZ growth process begins with the melting of polysilicon nuggets and doping element or alloy contained in a silica crucible. Temperature adjustments and stabilization follow the melt-down such that the melt surface temperature is slightly supercooled. A crystal seed ($\sim$12 mm in diameter) attached to the end of a steel cable is dipped into the melt. If the crystal-melt contact forms a smooth meniscus, the "necking" process based on Dash's [16] technique may begin. The reason for necking is as follows: the growth of dislocation-free silicon has to be "extended" from a dislocation-free silicon crystal. Although the seed crystals used in CZ are usually dislocation-free, upon contacting the melt, dislocations are generated due to thermal shock. These dislocations must be eliminated before the full diameter crystal growth can begin. Necking is a procedure to outgrow dislocations. During necking, the seed crystal is gradually decreased in diameter by increasing the pull speed and appropriate temperature adjustment. The goal is to reach a steady state neck growth condition with a neck diameter of 3–4 mm and a pull rate of 4–6 mm/min. The dislocation-free crystal is usually achieved after a few centimeters of growth. Two phenomena may occur during the necking. When the seed diameter is large in the early stage of necking, the thermal stress cause dislocation movements by glide, cross slip and climb mechanisms since the temperature is above the "plastic" temperature ($>900^{\circ}$C). In the growth of $\langle 111 \rangle$ and $\langle 100 \rangle$ crystals, the growth axes are oblique or perpendicular to $\{111\}$ slip planes. Therefore, the dislocations can glide out of the crystal surface after sufficient time. For $\langle 110 \rangle$ growth, because $\langle 110 \rangle$ is contained in a $\{111\}$ plane, the

seed has to be oriented a few degrees off the pulling axis towards the direction perpendicular to the (111) plane, to facilitate dislocation elimination. When the neck is 3–4 mm in diameter under high speed pulling, the stress is relatively small, causing slow or no movement of the prevailing dislocations. When dislocation movements are slower than the advancing solid–liquid interface, dislocation-free growth is obtained. Typically, the dislocation-free status is accompanied by the growth of strong ridges (⟨100⟩ crystals) or "flats" (⟨111⟩ crystals) on the crystal's symmetry positions, which are actually due to prominent {111} facet growth at these positions [17].

Once the dislocation-free growth is achieved through necking, the diameter may be expanded by "shoulder" growth until it reaches the desired diameter. The body growth is under automatic diameter control (ADC), in which the pull rate is slaved by optically monitoring crystal diameter variations. The ADC is also assisted by minor temperature adjustments slaved by the long-term pull rate changes. The crystal growth process is terminated by a gradual decrease from full diameter to zero in a low pull speed in order to minimize the thermal stress by diameter change and associated slip generations. After the heater power is off, the crystal usually stays in the grower for a period of time for cooling, before it is removed from the grower. The total dwell time and temperatures the crystal experiences during the crystal growth and cool down constitute the so-called "thermal history" of the silicon crystal. The thermal history of CZ silicon determines the state of nuclei for oxygen precipitation, which is unique to that crystal growth process and is an important consideration, as is oxygen concentration, in oxygen precipitation kinetics. The heart of a CZ puller is the hot-zone. The design of heater and heat shields, for example, determines the radial and vertical thermal gradients in the melt. These thermal characteristics are intimately related to the growth characteristics, such as interface shape and related thermal stress generation [18]. Although this concept is fundamental to large-diameter silicon growth, little is known about the correlation between the crystal growth system, hot-zone design factors and growth characteristics. For a given growing system, an empirically observed optimum growth condition is usually obtained by modifying the hot-zone components and/or by varying the growth parameters by trial and error. Hot-zone thermal characteristics impact many aspects of silicon crystal properties, including oxygen incorporation behavior and the thermal history of the grown crystal. Since hot-zone design varies from one crystal grower to another, the crystals grown from different growers are expected to differ in these properties. Thermal convection and forced convection conditions are also important controlling factors for CZ crystal properties and will be discussed in Section 3.3.

3.3 Characteristics of Czochralski Silicon Growth

3.3.1 Thermal Characteristics for Dislocation-Free Growth

In the growth of large diameter silicon crystals, the thermal characteristics of the growing system are the most important consideration. The structural morphology, perfection and other properties of large-diameter crystals are determined by the nature of the heat flow balance near the melt–crystal interface. Figure 3.3 shows the heat flow pattern in a CZ growing system. There is a balance of heat flow about the interface during the steady state growth, i.e., $Q_C + Q_R = Q_L + Q_T$. The interface shape and neighboring isotherms in the growing crystal are controlled by the crystal cooling conditions and melt radial temperature gradient, as well as by the solidification rate of the crystal. The crystal cools by radiation and gas convection at the crystal surface, while the melt radial temperature gradient is determined by the melt aspect ratio and heater geometry. When the crystal diameter is small, the crystal may be considered as a one-dimensional conductor. The surface cooling rate would, at best, affect the growth rate. As the diameter increases, the shape of the interface and isotherms in the crystal becomes highly dependent on the surface cooling rate, the melt radial temperature gradient and the growth rate. Excessive surface cooling, a shallow radial temperature gradient, or a high growth rate can result in highly concave interfaces and nearby isotherms. This configuration corresponds to a high thermal stress condition [19], which can generate slip dislocations when the thermal stress exceeds the critical resolved shear stress [20]. This usually occurs in the region above the growing interface where the temperature is in the plastic

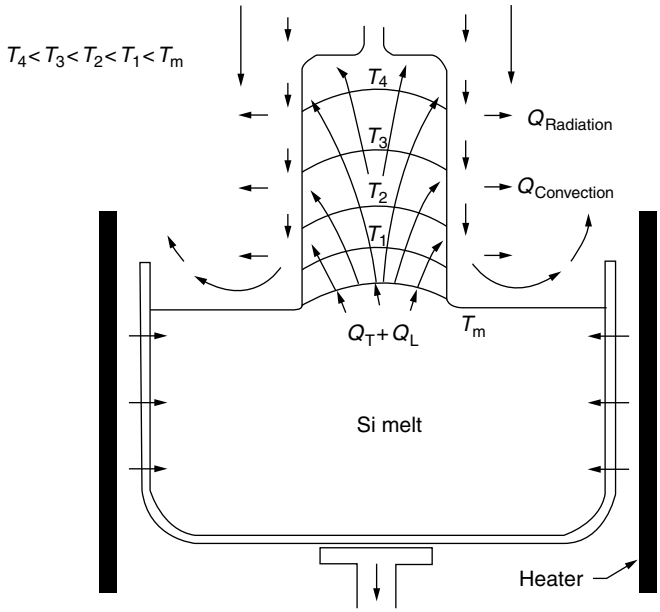


FIGURE 3.3 Schematic heat flow in a Czochralski (CZ) growth arrangement; Q_L is latent heat and Q_T is heat transfer to the crystal. (From Lin, W. and Benson, K. E., *Annual Review of Materials Science*, 17, 273, 1987. Reproduced with permission from Annual Reviews.)

deformation range ($>900°C$). Once dislocations are generated, they are propagated toward the interface and continued growth will not be dislocation-free. More often than not, through dislocation multiplication, the growing crystal eventually turns into polycrystalline silicon. Thermal stress-induced perfection loss is usually accompanied by a highly concave interface. Figure 3.4 is a cross-section of a 150-mm diameter silicon crystal grown under unfavorable thermal conditions. The growth experiments show that increasing the melt radial temperature gradient reduces the curvature of the interface and thus the thermal stress. The thermal stress in a CZ crystal can be modeled and the distribution of the total resolved shear stress on each slip system can be calculated, as was done by Jordan et al. [21]. Their model predicts slip dislocations, which are typically detected in poor-quality crystals.

The melt radial thermal gradient also has a profound impact on the crystal shape of large diameter crystals. For example, during the growth of ⟨111⟩ crystals, flat regions develop on the crystal periphery in the ⟨211⟩ directions. That is, the lateral growth is retarded in the ⟨211⟩ directions (see Figure 3.5a). The flats shown in Figure 3.5a result from an excessive supercooling, which is caused by a small radial temperature gradient in the melt or a high convective heat loss at the meniscus. In severe cases, the cross-section tends to be triangular. A steeper thermal gradient has been found to minimize such preferential growth.

It should be noted that the aforementioned shape problem due to flat formation is common in lightly doped ⟨111⟩ silicon growth. In the growth of heavily doped $N^+\langle 111\rangle$ crystals (such as those doped with antimony in the concentration range of $10^{18}/cm^3$), the situation is quite different. The preferred growth in this case is in the ⟨211⟩ direction and facet growth is retarded. The result is a cam-shaped crystal cross-section, as shown schematically in Figure 3.5b. A theoretical model that predicts the effect of dopant type and concentration on silicon growth habit has not been developed. The radial thermal gradient also affects the shape of ⟨100⟩ crystals. Small gradients can cause preferred lateral growth in the ⟨110⟩ directions, resulting in a rectangular or square cross-section (Figure 3.5c). The preferential growth is very sensitive to the thermal conditions and the shape of the crystals can be used as a measure of the thermal conditions of a growing system. This is especially true for lightly doped ⟨111⟩ crystals.

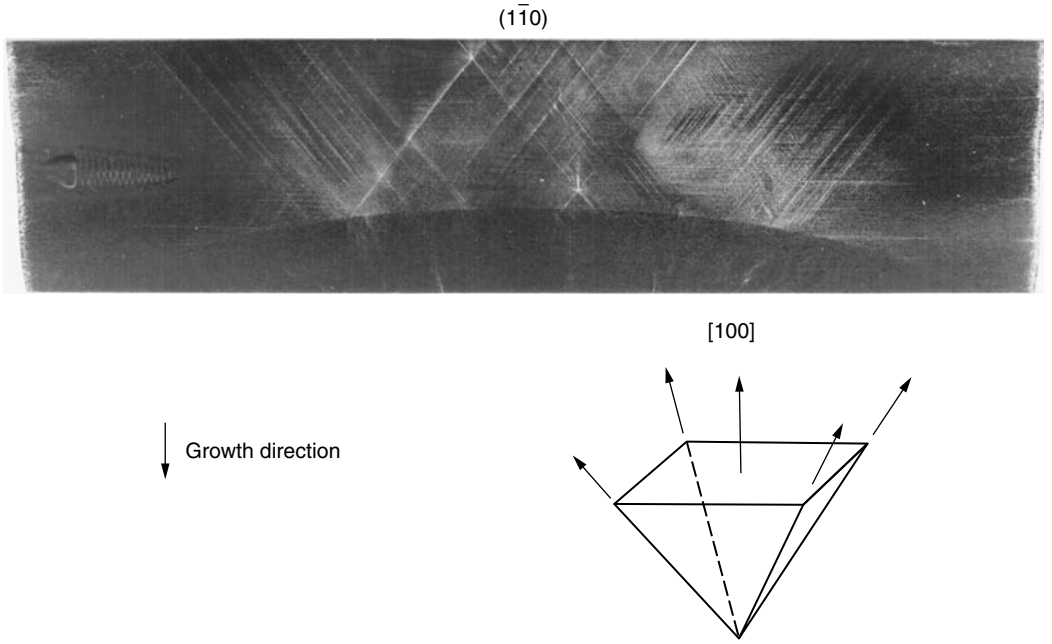


FIGURE 3.4 Etched longitudinal cross-section of a silicon crystal of 150 mm in diameter showing stress-generated slip dislocations from an instantaneous interface (curved demarcation) and its propagation above the interface. (From Lin, W. and Benson, K. E., *Annual Review of Materials Science*, 17, 273, 1987. Reproduced with permission from Annual Reviews.)

In growing crystals of large diameter, the main concern is the thermal stress build up due to a massive latent heat of solidification at the interface and longer path for heat dissipation during the crystal growth. To maintain the heat balance at the interface for dislocation-free growth discussed above, the growth rate has to be reduced as the crystal diameter is increased. This trend has been observed in the development of 300- and 400-mm-diameter crystal growth, as will be discussed in a later section. From thermal stress considerations, there will be a practical upper limit for the diameter in CZ silicon crystal growth.

3.3.2 Impurity Incorporation

3.3.2.1 Axial Dopant Distribution

In the growth of silicon crystals from large melts using ADC mechanisms; involving pull rate and/or temperature changes slaved to optical diameter measurements, the axial dopant distributions follow the normal freezing behavior [22],

$$C_s(x) = C_0 k(1-x)^{1-k}, \tag{3.2}$$

where C_s, k, and x are crystal dopant concentration, dopant segregation coefficient and fraction of melt solidified, respectively. C_0 is the initial dopant concentration of the melt. Complete mixing takes place in the melt by vigorous thermal convection and the segregation coefficient of the dopant assumes the equilibrium value k_0 in most instances. In the reduced pressure growth of heavily doped silicon involving dopants with high vapor pressure, such as antimony, the k value can deviate significantly from the equilibrium value (k assumes a value greater than k_0 but less than unity). Due to dopant segregation, there is a spread in dopant concentration along a CZ crystal. The degree of the spread depends on the k

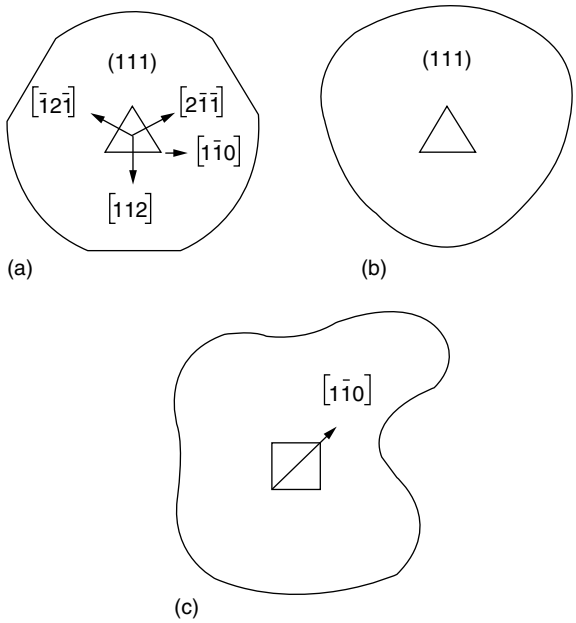


FIGURE 3.5 Cross-section shapes of crystals grown under unfavorable thermal conditions: (a) lightly doped *p*-type (111); (b) heavily Sb doped (111); (c) lightly doped (1$\bar{1}$0). (From Lin, W. and Benson, K. E., *Annual Review of Materials Science*, 17, 273, 1987. Reproduced with permission from Annual Reviews.)

value of the dopant. Thus, the smaller the *k* value, the larger the spread in concentration. The segregation effect of the dopants often limits the "yield" of the CZ silicon crystals. The non-standard CZ methods, such as double crucible method [17,23] and continuous growth method [24] have been used to eliminate the effect due to segregation (see Figure 3.6). The axial microscopic uniformity is controlled by the microscopic growth rate. Thermal convection, thermal asymmetry, and pull-rate fluctuations are sources of microscopic growth fluctuations. The nature of the microinhomogeneity of impurity in CZ crystals will be discussed, following the discussion of melt convection flows and effective segregation coefficient.

3.3.2.2 "Unintended Dopants"

Impurities in the crucible material or the vapor above the melt can be incorporated into the melt and the silicon during crystal growth. Oxygen and carbon are the major impurities incorporated into CZ silicon. Their concentrations in CZ crystals are on the order of 10^{18} and 10^{16} atoms/cm^3 for oxygen and carbon, respectively. The silica crucible is an infinite source for oxygen. Molten silicon dissolves the silica and absorbs oxygen. Unlike normal dopants, oxygen in the silicon melt is a dynamic system and the oxygen distribution is not homogeneous in the melt. The oxygen concentration incorporated into the crystal is a result of complex interactions between the crucible dissolution rate and the nature of the melt flow (which determines how the oxygen-rich melt is transported). Therefore, the axial oxygen profile of a silicon crystal is not the result of normal freezing behavior and the concentration profile can vary widely depending on the grower thermal characteristics and the growth parameters used.

The source of the carbon is the graphite material making up the hot-zone of the crystal grower. The silicon monoxide evaporated from the melt surface interacts with hot graphite components and is reduced to carbon monoxide before re-entering the melt following:

$$\mathrm{SiO} + 2\mathrm{C} \rightarrow \mathrm{SiC} + \mathrm{CO} \tag{3.3}$$

The introduction of CO into the melt is a continuous process. The incorporation from the vapor phase and residual carbon content in the starting polysilicon ($<$0.2 ppma) results in a normal freezing-type

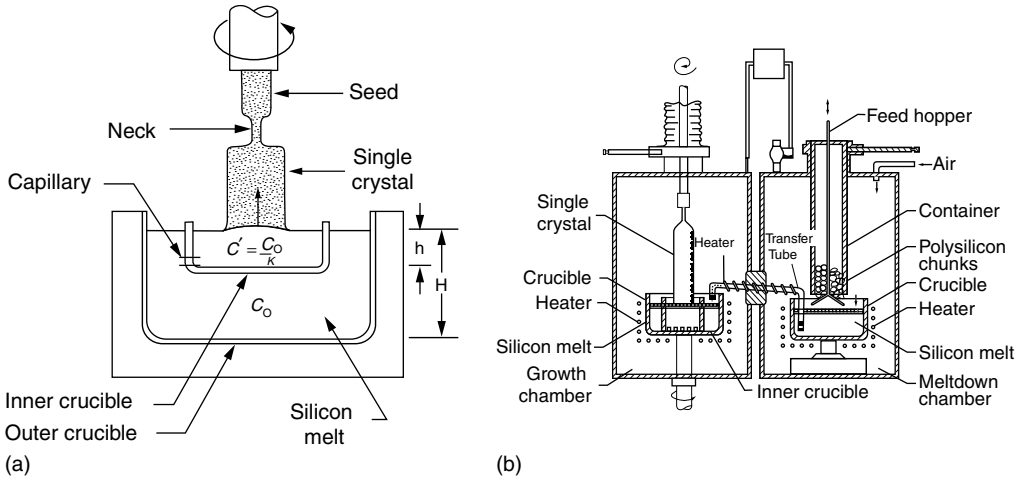


FIGURE 3.6 (a) Schematic representation of a double crucible growth arrangement (b) Continuous liquid-feed Czochralski growth furnace. (From Benson, K. E., Lin, W., and Martin, E. P., *Semiconductor Silicon 1981*, eds, Huff, H. R., Kregler, R. J., and Takeishi, Y., Electrochemical Society, Princeton, NJ, 1981, 33 and Lorenzini, R. E., Iwata, A., and Lorenz, K., U. S. Patent 4,036,595, 1977. Reproduced with permission from Electrochemical Society.)

carbon concentration profile in CZ silicon (carbon has a small segregation coefficient, $k_0 = 0.07$). Carbon in silicon has not been shown to affect IC characteristics. Carbon concentration of up to 4 ppma did not directly change dynamic random access memory (DRAM) performance or yield [26]. However, carbon in silicon has been observed to enhance oxygen precipitation. To utilize this property for oxygen precipitation enhancement, one must grow silicon with controlled carbon concentration and uniformity. Such silicon growth process has not been reported. For ULSI applications, it is a common practice to keep carbon at minimum levels (<0.5 ppma).

Besides carbon and oxygen, metal atoms can also be incorporated into the growing crystal. Metallic elements, especially transition metals are among the most undesirable contaminants in silicon material for IC processing due to its role as lifetime killers. They are also fast diffusers in silicon at elevated temperatures and even at 700°C. Fortunately, due to the small values of their segregation coefficients (10^{-4}–10^{-6}) [27], no significant metal can be grown into silicon crystals when the melt is contaminated with metals [28]. However, transition metals transported in the vapor phase can condense and diffuse through the silicon crystal surface at high temperatures and rapidly into the bulk. The diffusion of metal contaminants is most "effective" at the grown crystal surface above the melt. The metal contaminants can be evaporated from an overheated metal or alloy surface of grower components. An example is the stainless steel shaft/cable for crystal pull and rotation, which is usually not water-cooled. Overheating of these parts occur when they are too close to the hot-zone. The use of an extended seed chuck (seed holding device attached to the bottom of the steel shaft/cable, made of ceramic materials or graphite) would minimize the overheating of the steel shaft/cable. Contaminated graphite for fabrication of hot-zone parts, such as heater, suscepter, and heat shields is another possible source of metal contamination. Metal evaporated from the graphite parts can contaminate the grown silicon via solid phase diffusion [29].

3.3.2.3 Macroscopic and Microscopic Inhomogeneity

3.3.2.3.1 Effective Segregation Coefficient, k_{eff} and Interfacial Boundary Layer

The degree of incorporation of an impurity at the freezing front is controlled by its segregation coefficient. When the solidification rate is low or the impurity concentration is dilute, the segregation

coefficient assumes a value very close or identical to the equilibrium value. In CZ silicon growth with finite or higher growth rate, the dopant ($k_0 < 1$) incorporation behavior was found to result in a k value which deviated from the equilibrium value, for example, the k was observed to increase with increasing growth rates [30]. This behavior is due to the finite diffusivity of the dopant in the crystal that cannot equalize the concentration change at the solidification front quickly. Because the concentration of impurities is higher at the solidification front, a larger number of impurity atoms are incorporated into the crystal than would be expected from the concentration in the bulk melt outside the diffusion-controlled layer. This results in a higher degree of impurity incorporation and an effective segregation coefficient k_{eff}, which is larger than the equilibrium value, k_0. Burton, Prim, and Schlicter (BPS) [31] extended the plain rotating disc treatment [32] for the diffusion boundary layer, for crystal growth, and showed that the degree of incorporation, or the effective segregation coefficient of an impurity, is related to the diffusion boundary layer thickness, δ, and growth rate f as:

$$\frac{k_0}{k_{\text{eff}} = k_0 + (1 - k_0)\exp(-\delta f/D),} \tag{3.4}$$

where D and k_0 are the diffusion coefficient and equilibrium segregation coefficient of the impurity, respectively. The thickness of the diffusion boundary layer δ depends on the diffusivity of the impurity, kinematic viscosity υ of the melt and crystal rotation rate ω as:

$$\delta = 1.6 D^{1/3} \upsilon^{1/6} \omega^{-1/2}. \tag{3.5}$$

From Expression 3.4 and Expression 3.5, it is seen that the instantaneous incorporation rate is microscopic growth rate-controlled. The growth rate fluctuations are the source of microscopic inhomogeneity in the crystal. On the other hand, at any given interface, the uniformity of the boundary layer thickness across the interface controls radial uniformity. The CZ crystal's radial and microscopic impurity uniformity will be discussed later in terms of the BPS equation.

3.3.2.3.2 Convection Flows in Czochralski Melt

The major convective fluid flows in the CZ melt that impact the growth are thermal convection flows due to the existence of a non-vertical temperature gradient, and forced convection induced by the crystal and the crucible rotation as shown in Figure 3.7 (other convection flows are caused by crystal pulling and surface tension). The basic thermal convection flow generated in a melt can be symmetrical or asymmetrical, with the hot melt rising along the crucible wall and descending at the crucible center as is shown in Figure 3.7a. The basic flow pattern is determined by the crucible geometry, aspect ratio (height to diameter ratio of the melt), and thermal boundary conditions. The driving force is given by the dimensionless Grashof number N_{Gr}:

$$N_{Gr} = g\alpha\, \Delta T R^3/\upsilon^2, \tag{3.6}$$

where g is the acceleration due to gravity, α is the melt thermal expansion coefficient, ΔT is the temperature difference across the crucible diameter R, and υ the kinematic viscosity of the melt. From this expression, it is clear that, as the crystal grower is scaled up in charge size (so is the crucible diameter), a stronger thermal convection is generated due to the direct contribution of the third power of R and increasing temperature gradient in the melt. Melt turbulence associated with strong thermal convection cause temperature fluctuations in the melt and contribute to microinhomogeneity in dopant and oxygen incorporation.

It is a common practice in CZ growth to use crystal rotation to override the deleterious effects of thermal convection discussed above. The effects of crystal rotation are twofold. Similar to a rotation disc, the crystal rotation induces a uniform hydrodynamic boundary layer across the interface. Burton, Prim, and Schlicter [31] extended the plain rotating disc treatment for the diffusion boundary layer for crystal

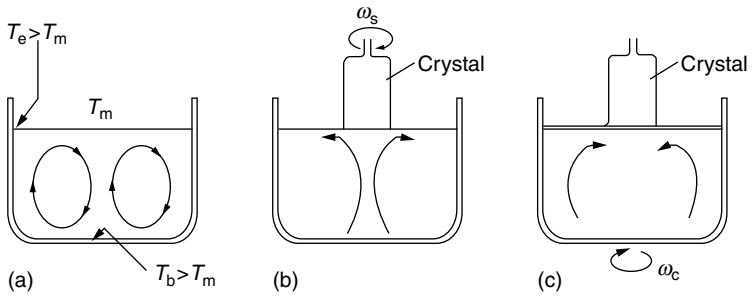


FIGURE 3.7 Convection patterns in a Czochralski melt due to (a) thermal convection, (b) crystal rotation and (c) crucible rotation. (From Lin, W., *Oxygen in Silicon*, ed. Shimura, F., Academic Press, 1994, chap. 2. Reproduced with permission from Elsevier.)

growth, in Equation 3.5. Secondly, the rotating crystal draws a uniform flow from the central region of the melt, perpendicular to the interface over the radius and spins it outward radially near the surface, Figure 3.7b. As a result, the convection flow by the crystal rotation has the effect to counter and reduce the thermal convection flow, and the convection flow induced by the crucible rotation (has same general flow pattern as thermal convection. See Figure 3.7c). The magnitude of the flow induced by the crystal rotation is characterized by the dimensionless Reynolds number N_{Re}:

$$N_{\mathrm{Re}} = \omega r^2/v, \tag{3.7}$$

where ω is the crystal rotation rate and r is the crystal radius.

While forced convection, i.e., the crystal rotation, has the effect of overriding the harmful thermal convection, the net effect depends on the relative magnitude of the two components. The relative effect of the two components may be expressed by the ratio, N_{Re}^2/N_{Gr}. If $N_{\mathrm{Re}}^2 > N_{Gr}$, the crystal rotation will effectively isolate the segregation process at the growth interface from the thermal convection in the melt [33]. Therefore, small melt and high crystal rotation would suppress the effect of thermal convection.

3.3.2.3.3 Macroscopic Radial Impurity Uniformity

In the silicon growth from a homogeneous melt, the radial uniformity is controlled by the uniformity of boundary layer thickness across the interface. In Equation 3.4, if the growth rate is assumed to be uniform across the interface, then the radial uniformity of the impurity is determined by the variation in boundary layer thickness. It may be shown [34], in the case of dopant in silicon, that the rate of change in k_{eff} with respect to the variation in diffusion boundary layer thickness is very sensitive in the range of layer thickness encountered in the silicon CZ growth. The effect is greater for n-type than for the p-type dopant in silicon. The relevance of the melt convection condition to the radial impurity incorporation is the following. In the absence of thermal convection, the result of the crystal rotation would be a uniform diffusion boundary layer over the crystal radius at the interface. Hence, if the crystal rotation is the only source of convection, no radial segregation will result from the fluid flow effect and impurity incorporation is uniform across the crystal radius. In real crystal growth, however, the thermal convection flows form a general pattern in the crucible, as rising streamlines along the crucible wall, which fall in the center following a gradually curved path (see Figure 3.7a). This path results in a stagnation point near the center of the crystal interface. Therefore, when thermal convection is strong compared to the forced convection, the diffusion boundary layer thickness at the outer region of the interface may be reduced by the thermal convection velocity, while the effect is small in the region near the center of the interface. Thus, thermal convection flow causes a radial variation in a boundary layer thickness. If the segregation coefficient of the dopant involved, k_0, is smaller than unity, one finds a

significant radial variation in impurity incorporation; it is high in the center and low in the outer region of the crystal. The greater the k value deviates from unity, the greater the radial variations. By the same principle, increased crucible rotation can cause variations in the boundary layer thickness and radial gradient in incorporation. As discussed above, harmful thermal convection effects can be suppressed by increased crystal rotation and small melt growth. Lin and co-workers [17,23] have demonstrated such effects by using a double crucible arrangement (see Figure 3.6a). The use of a smaller diameter and a low aspect ratio of inner crucible reduced the effect of thermal convection. The improvements on radial dopant uniformity and random dopant concentration fluctuations, and characteristic of the thermal convection, were observed for As, P, Sb, and oxygen for which the segregation coefficients deviated significantly from unity.

3.3.2.3.4 Microscopic Inhomogeneity

In general, the microscopic inhomogeneity of impurity in CZ silicon crystals is a result of growth rate fluctuations and impurity segregation during crystal growth. The growth rate fluctuations cause variations in the impurity incorporation levels. The lattice strain associated with local impurity concentration variations give rise to the so-called "striations," which are revealed by chemical etching or x-ray topography. Severe microscopic dopant inhomogeneity corresponds to a large variation in carrier concentration and is not desirable in silicon materials used for device fabrication, especially when such variation is comparable to the device feature size. This is an important consideration in VLSI/ULSI fabrication. Large oxygen striations can result in preferential precipitation, often observed as concentric ring patterns in etched wafers following thermal processing. In CZ silicon growth, there are several sources of microscopic growth rate variations which are discussed below.

3.3.2.3.4.1 Non-Centrosymmetric Thermal Distribution in the Silicon Melt

In large-melt silicon growth, finite thermal asymmetry exists about the center of the melt. During crystal growth, as the crystal is rotated about the growth axis, the interface experiences slightly different temperatures at different positions in the melt. Therefore, the growth rate of a given crystal element parallel to the crystal axis, fluctuates periodically, as illustrated schematically in Figure 3.8. In general, the fluctuation is most pronounced in the crystal elements furthest from the crystal center. The periodicity of the fluctuation is determined by the average growth rate, f, and crystal rotation rate ω, as f/ω. The variations in the impurity incorporation level that correspond to the growth rate variations are commonly referred to as "rotational striations." When the equilibrium segregation coefficient (k_0) of the element involved is less than unity, the fluctuation in the impurity incorporation level is in phase with the growth rate fluctuation. If $k_0 > 1$, the fluctuations will be out of phase (see Figure 3.8). The relationship may be realized readily by examining the Equation 3.4.

3.3.2.3.4.2 Thermal Convection-Related Temperature Fluctuations

Growth rate fluctuations caused by thermal convection related temperature variations are mostly random in nature and etched striations are characteristically aperiodic. When thermal convection is significant, the microstriations bear the signature of high-frequency fluctuations in the order of tens of hertz.

3.3.2.3.4.3 Automatic Diameter Control-Induced Perturbation

Growth-rate fluctuation can be further perturbed by the ADC commonly employed in silicon crystal growth. The crystal pull-rate is slaved by optically monitored crystal diameter variations, in order to maintain a preset diameter. The pull-rate adjustments are both "instantaneous" (a few seconds) and "long term" (minutes). The long term pull-rate adjustment determines the average growth rate. The instantaneous pull-rate adjustment imposes modifications on the microscopic growth rate fluctuations resulting from thermal asymmetry and thermal convection. The net effect is to smear the periodic nature of the impurity fluctuation resulting from the melt thermal asymmetry.

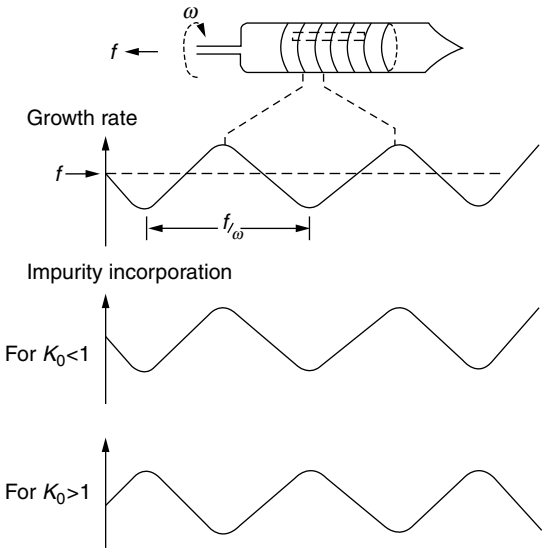


FIGURE 3.8 Schematic illustration of growth rate fluctuation and its relationship with microscopic impurity fluctuations in CZ crystal. Here, f and ω are growth rate and crystal rotation rate, respectively. (From Lin, W. and Stavola, M., *J. Electrochem. Soc.*, 132, 1412, 1985. Reproduced with permission from Electrochemical Society.)

3.3.2.4 Oxygen Incorporation and Segregation in Czochralski Silicon Growth

3.3.2.4.1 Incorporation Mechanism

Unlike most intended dopants in silicon CZ crystal growth, which show "normal freezing" behavior under normal growth conditions, oxygen is an unintended dopant that enters the silicon melt continuously by dissolving the silica crucible. The incorporation behavior of oxygen into silicon is the result of complex interplay among crucible dissolution, surface evaporation, thermal convection and forced convection, as shown in Figure 3.9. In a side-heated CZ hot-zone, the dissolution of the silica crucible is the highest at the side wall of the crucible. The silicon melt dissolves SiO_2 and absorbs its oxygen. The oxygen-rich silicon melt would rise along the crucible wall, following the thermal convection flow pattern to near the melt surface, and then to the melt center where it is drawn toward the crystal for incorporation by the forced convection induced by the crystal rotation. When the oxygen-rich melt is near the surface, a great portion of the oxygen is evaporated through the free surface. The oxygen concentration incorporated into the growing crystal, therefore, is proportional to the oxygen concentration in the melt adjacent to the growing crystal. During steady-state growth, there exists a dynamic equilibrium of oxygen between the four controlling factors in the system. Due to the difference in thermal characteristics of different hot-zone designs, the oxygen incorporation behavior varies from one grower design to another. The major oxygen-controlling factors are discussed in the following.

3.3.2.4.1.1 Effect of Surface Evaporation and Crucible Dissolution

In the absence of thermal and forced convection (a hypothetical case), the melt oxygen concentration is proportional to the ratio of melt-crucible contact area to the available free surface area. In such a case, the transport of oxygen would depend entirely on diffusion. This ratio constantly decreases as the aspect ratio is reduced during the growth. The variation of this ratio is the basic characteristic of the axial oxygen profile of CZ silicon, in which the concentration is observed to gradually decrease from the seed end toward the tang end (for example, see Figure 3.15). However, the dependence of concentration on this geometric ratio is influenced by the crucible dissolution rate and ambient pressure. The dissolution rate

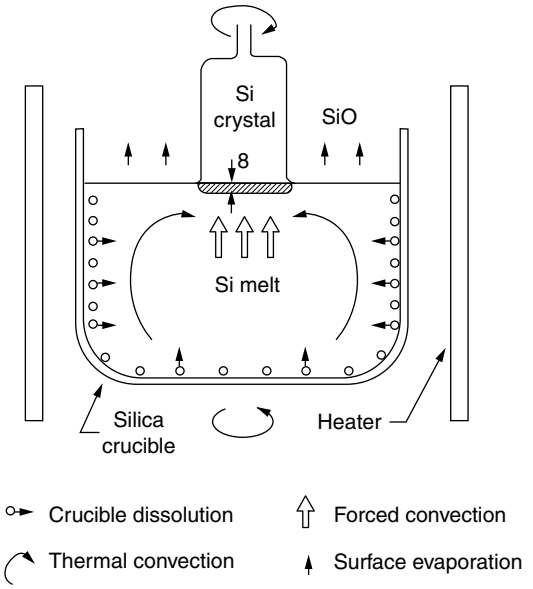


FIGURE 3.9 Schematic of a silicon Czochralski growth system showing relationship among oxygen-controlling factors. (From Benson, K. E., Lin W., and Martin, E. P., *Semiconductor Silicon 1981*, eds. Huff, H. R., Kregler, R. J., and Takeishi, Y., Electrochemical Society, Princeton, NJ, 1981, 33. Reproduced with permission from Electrochemical Society.)

of the silica is material density and temperature dependent. Therefore, using crucible with an inner wall made of "porous silica," for example, would enhance the dissolution rate. Increasing the melt-crucible contact area by using corrugated inner surface will obviously serve similar purpose. Mathematical modeling of the observed oxygen incorporation behavior based on dissolution and evaporation, as carried out by Carlberg et al. [35], is merely a first-order approximation. In actual silicon growth, oxygen incorporation and uniformity are greatly affected by thermal convection and forced convection. No modeling effort thus far accurately takes these factors into account.

A study of oxygen evaporation from the melt surface was made using "shoulder" growth of a 300-mm-diameter crystal from a 350-mm-diameter crucible. In this experiment, the effect of the free melt surface evaporation on the oxygen level during crystal crown growth was examined. Figure 3.10 plots oxygen concentration variation in the grown silicon, as the melt surface is covered by the growing "shoulder." The relationship between oxygen concentration and available surface is not linear. When the crystal shoulder diameter is small ($<$125 mm), the oxygen evaporation (and therefore the oxygen concentration of the growing crystal) is very sensitive to the diameter change. When the majority of the surface is covered with the crystal, the evaporation seems to remain nearly constant. By extrapolating the curve in Figure 3.10, the oxygen concentration corresponding to a very small and to a maximum size of crystal (when the melt is fully covered) may be obtained. The maximum oxygen evaporation accounts for about 30% of the available dissolved oxygen. This value is much lower than generally conceived, that over 90% of the dissolved oxygen from crucible is evaporated from the melt surface [36]. More discussion of the oxygen incorporation behavior on the 300-mm-diameter crystal growth will be made in a later section.

Lin and Hill [17] studied the effect of ambient pressure on the oxygen distribution in the melt and grown crystal, via crystal growth experiments. Based on the experimental evidence, physical models of oxygen distributions in large silicon melts were proposed and are shown schematically in Figure 3.11. It is shown that the thermal convection is the main oxygen transport mechanism when the melt aspect ratio is high in a large melt growth. The oxygen-rich flow conforms with the suggested thermal convection

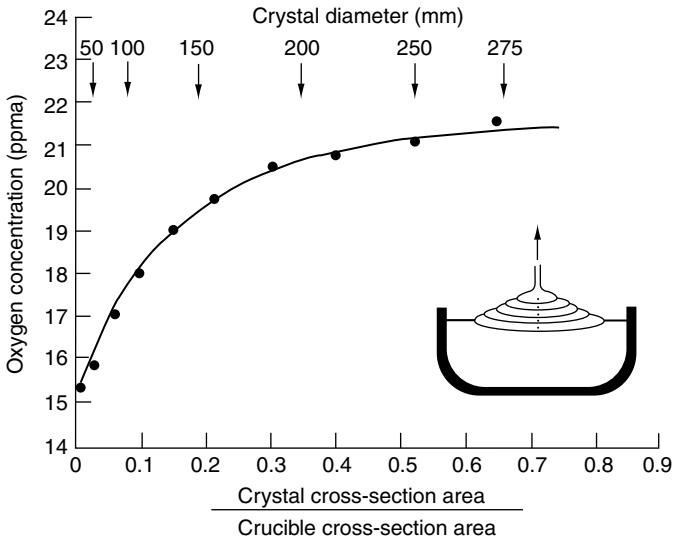


FIGURE 3.10 Oxygen concentration measured as a function of fraction of melt surface being covered by the growing crystal during crown growth. The crystal crown accounts for less than 10% of the initial charge. (From Lin, W. and Benson, K. E., *Annual Review of Materials Science*, 17, 273, 1987. Reproduced with permission from Annual Reviews.)

pattern shown in Figure 3.7a. Experimental results also showed that under atmospheric pressure, the oxygen distribution is not uniform near the melt center, where a stagnant region of low oxygen concentration exists. The nature of this non-uniformity is displayed in crystals' radial oxygen profiles near the seed end, in Figure 3.12a. However, as the crystal growth progresses, the forced convection due to crystal rotation modulates the oxygen distribution and results in the local mixing, and the gross non-uniformity near the stagnant region is largely diminished. On the other hand, under the reduced pressure, Figure 3.12b, the oxygen distribution in the melt is more uniform near the surface of the melt. The grown crystal possesses better radial oxygen uniformity. When thermal convection is not significant, as in the case of a low aspect ratio configuration, the melt oxygen is both diffusion rate-dependent and crucible dissolution rate-dependent. The melt tends to be uniform in oxygen. Both enhanced forced convection and ambient pressure have little effect on the oxygen uniformity.

3.3.2.4.1.2 Effect of Crystal and Crucible Rotations

For a given system, i.e., fixed starting melt geometry and hot-zone thermal distribution, etc., the parameters that can significantly alter the oxygen incorporation are crucible and crystal rotations, and growth rate variations. The effect of crystal/crucible rotation on the fluid flow patterns were studied in the past by simulation using fluid of similar viscosity as that of silicon melt at room temperature [37]. Figure 3.13 shows the fluid flow patterns due to various combinations of crystal and crucible rotations. In the real crystal growth, however, the flow patterns can be significantly altered by the presence of thermal convection. The results of the simulations provide very useful information on the effect of rotational parameters. Kakimoto et al. [38,39] observed thermal and forced convection flows of silicon melt directly during Czochralski growth using x-ray radiography with solid tracers, for various crystal and crucible rotation speeds. The effect of non-axial, symmetrical temperature distributions on the thermal convection flows was clearly observed. The suppression of thermal convection by crystal rotation-induced forced convection was also evidenced. One way to gain information on the flow properties of a growing system is to analyze grown crystals following parametric growing experiments. One finds that forced convection is effective in controlled oxygen incorporation.

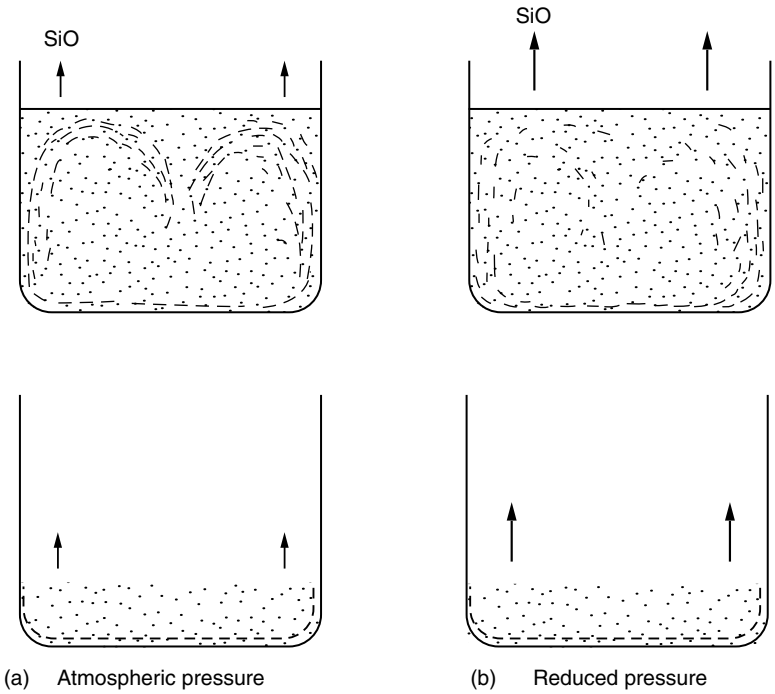


FIGURE 3.11 Schematic representations of oxygen distribution in silicon melt at high and low melt level configurations. The dots and lines represent oxygen concentration. (a) At atmospheric pressure. (b) At reduced pressure. (From Lin, W. and Hill, D. W., *Silicon Processing*, ASTM STP 804, 1983, 24. Reproduced with permission from ASTM.)

As discussed previously, crystal rotation rate determines the magnitude of the upward melt flow. This flow can serve as oxygen transport from crucible bottom to the growing interface. The net effect on the overall flow pattern and oxygen incorporation depends on its magnitude and rotational direction relative to the crucible rotation. Figure 3.14 shows an "uncommon" oxygen profile of silicon grown under high crystal rotation (30 rpm) with crucible in iso-rotation mode (2 rpm). At about 30% of melt solidified, the oxygen incorporation underwent a mode change and sharply increases to a very high concentration level ($\sim$25 ppma). This behavior indicates that as the melt aspect ratio is reduced during the crystal growth, at some transition point, the strong forced convection takes over as the dominant transport mechanism which draws oxygen-rich melt from the crucible bottom to the growing interface.

Crucible rotation develops radial pressure gradients, which enhance the thermal convection flow arising from non-vertical temperature gradient as shown in Figure 3.7c. Therefore, fast crucible rotation helps the transport of oxygen from near the crucible wall to the growing crystal and enhances incorporation. Figure 3.15 shows the effect of crucible rotation on the incorporation level. Fast melt flow also results in a thinner melt-crucible boundary diffusion layer, a condition that will enhance crucible dissolution. Figure 3.16 shows several axial oxygen concentration profiles of silicon grown with several combinations of crystal/crucible rotation rates, under both counter- and iso-rotation conditions in the same grower and a reduced pressure. These results show that the forced convection induced by crystal/crucible rotations has very significant effects on the melt flow pattern, even in the presence of thermal convection. The bulk of the incorporation behavior is consistent with, and can be interpreted from, the simulated flow patterns (see Figure 3.13).

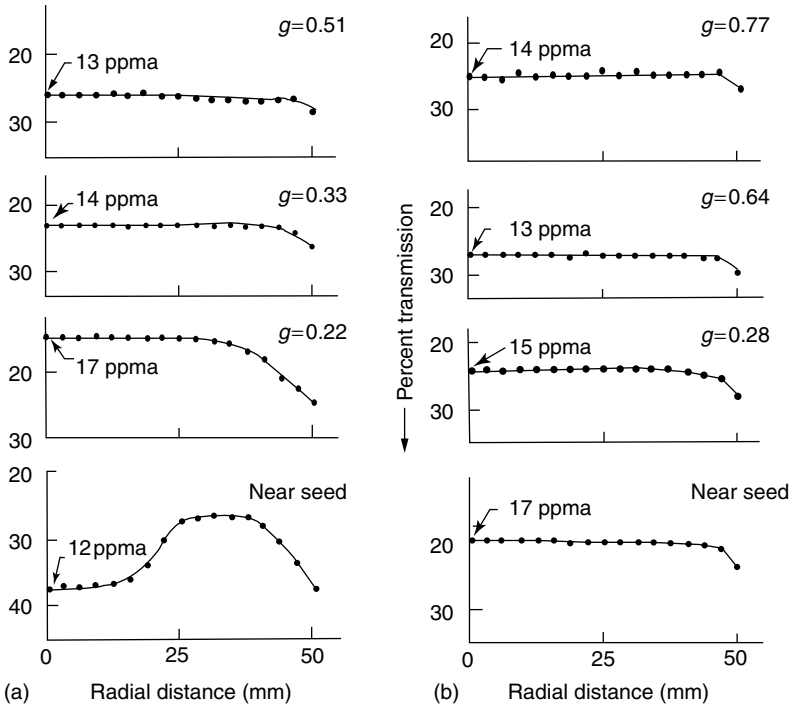


FIGURE 3.12 Radial profiles of 9 μm IR transmission at various stages: g=fraction solidified. (a) At atmospheric pressure. (b) At reduced pressure (18 torr). Oxygen concentration is computed in accordance with ASTM F121-80. (From Lin, W. and Hill, D. W., *Silicon Processing*, ASTM STP 804, 1983, 24. Reproduced with permission from ASTM.)

3.3.2.4.1.3 Incorporation in p^+ and n^+ silicon crystal growth

Degenerately doped n- and p-type silicon, in the concentration range of 10^{18}–10^{19} atoms/cm^3 are common substrate materials for n/n^+ and p/p^+ epitaxial structures for complementary metal–oxide–silicon (CMOS) ICs. Unlike the lightly doped silicon, oxygen precipitation behavior in degenerately doped silicon is drastically different depending on the conductivity type [41]. In this resistivity range (0.005–0.02 ohm–cm), oxygen precipitation is retarded in Sb-doped silicon while the kinetics are at their peak for p-type, boron-doped silicon. The possible sources of differences in precipitation kinetics in p^+ and n^+ were investigated. Oxygen diffusion mechanism was found not to be affected by the presence of heavy Sb or light boron [42]. The nucleation mechanisms for oxygen precipitates have been thought to be different in p^+ and n^+. Experimentally, the oxygen incorporation in n^+ and p^+ have been found to be different from the lightly doped silicon [43].

Figure 3.17 shows the axial oxygen distribution in 100 mm diameter p^+(0.005–0.01 ohm–cm) and n^+(0.02–0.08 ohm–cm) crystals as measured by SIMS, compared to the distribution band of $p-$ (8.20 ohm–cm) grown with the same conditions. It is seen that on the average, the p^+ crystals exhibit about 25% higher incorporation rate than n^+ crystals (with p^+ contents higher than p^-, while n^+ are lower). Other studies [44,45] also show the dependency of oxygen incorporation on doping level, but to different degrees. In considering the oxygen incorporation mechanism, it would not be surprised to find that the dependency obtained would vary somewhat depending on the melt size, melt aspect ratio, etc., used in the crystal growth experiments. Several possible mechanisms behind the dependency of oxygen incorporation on the dopant concentration have been suggested [46–48]. Some data indicate that the reduced oxygen incorporation into heavily Sb-doped silicon is due to Sb_2O_3 evaporation from the melt,

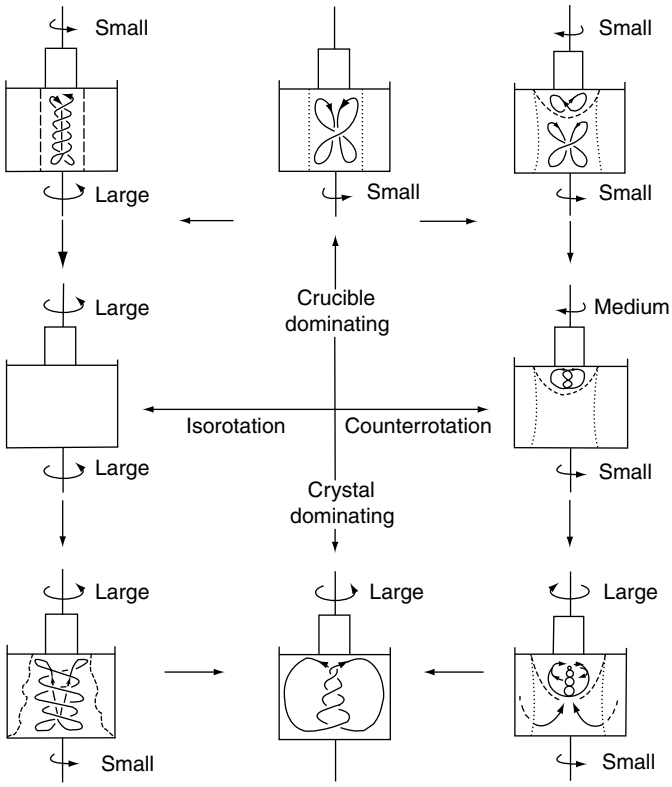


FIGURE 3.13 The variation of Czochralski flow patterns with relative directions and magnitudes of crystal and crucible rotations. (From Carruthers, J. R. and Nassau, K., *J. Appl. Phys.*, 39, 5205, 1968. Reproduced with permission from AIP.)

thus reducing oxygen concentration in the melt [46]. Others explain the lower oxygen incorporation observed in heavily Sb-doped silicon as due to accelerated SiO evaporation from the melt caused by simultaneous evaporation of elemental antimony [48]. In the case of p^+ crystal growth, the enhanced oxygen incorporation is speculated due to enhanced crucible dissolution by heavily boron-doped silicon melt. It is pointed out here that the difference in oxygen incorporation level cannot account for the difference observed in precipitation kinetics in p^+ and n^+. The difference in the nucleation mechanism probably plays a significant role.

3.3.2.4.2 Oxygen Segregation and Microscopic Inhomogeneity

Solute segregation during the solidification of a binary system is determined by the nature of its phase diagram near the solvent's melting temperature. The equilibrium segregation coefficient of the solute is a physical constant and is related to the slopes of the liquidus and solidus immediately adjacent to the melting temperature, above the primary phase of the equilibrium phase diagram as is shown in Figure 3.18. The equilibrium segregation coefficient, k_0 can be lesser or greater than 1 and can be deduced readily when the equilibrium phase diagram near the primary phase is established. In practice, when the solidification is not under equilibrium condition, the liquidus and solidus will shift in positions and segregation coefficient will deviate from k_0. Solute segregation phenomenon can be demonstrated by a controlled directional solidification of a binary alloy, in which the solute concentration along the direction of solidification will follow Equation 3.1. If the solution is in good mixing condition and solidification rate is small, the k value can approach k_0.

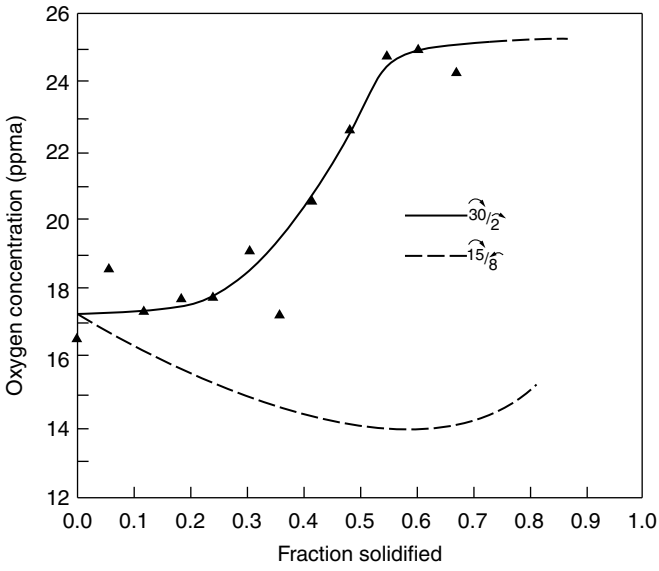


FIGURE 3.14 Axial oxygen profile of silicon grown with crystal rotation of 30 rpm and crucible rotation of 2 rpm (in iso-rotation). The dashed line represents profiles due to normal growth. (From Lin, W., *Oxygen in Silicon*, ed. Shimura, F., Academic Press, 1994, chap. 2. Reproduced with permission from Elsevier.)

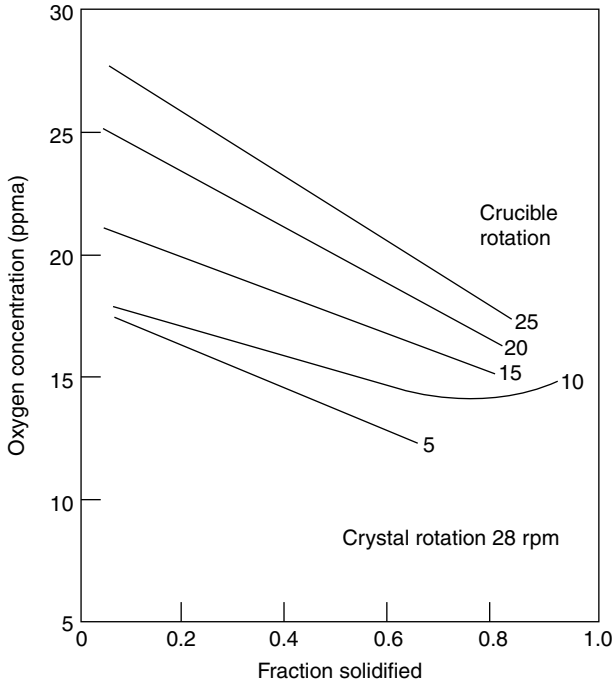


FIGURE 3.15 Oxygen profiles at various crucible rotation rates, with crystal rotation rate (in counter rotation) held at 28 rpm. (From Moody, J. W., *Semiconductor Silicon 1986*, eds. Huff, H. R., Abe T., and Kolbeson, B., Electrochemical Society, Pennington, NJ, 1986, 100. Reproduced with permission from Electrochemical Society.)

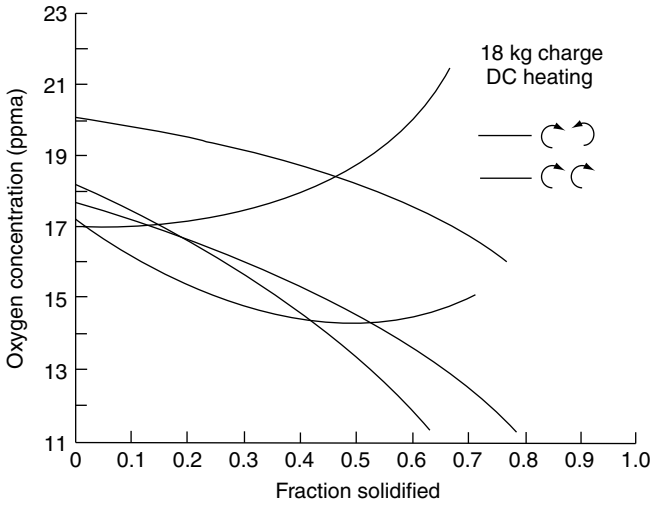


FIGURE 3.16 Axial oxygen profiles of silicon crystal grown with several combinations of crystal and crucible rotation rates. (From Lin, W. and Benson, K. E., *Annual Review of Materials Science*, 17, 273, 1987. Reproduced with permission from Annual Reviews.)

In the growth of CZ silicon, the dopant segregation behavior is similar to the solute in the directional solidification of a binary alloy and the dopant distribution along the grown crystal will follow Equation 3.1. If the dopant is non-volatile, such as boron or phosphorus, and melt is in complete mixing, the k value approaches the equilibrium value. However, in the case of oxygen, the incorporation behavior is quite different from the dopant element. The crucible dissolution rate, the oxygen transport mechanisms

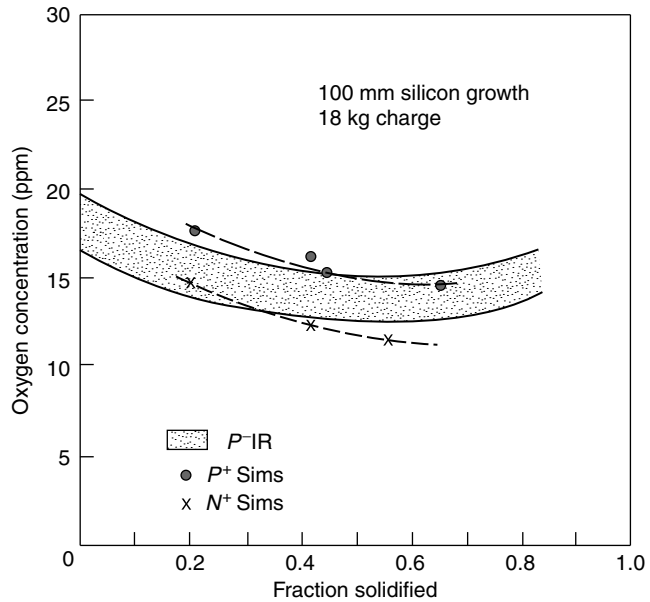


FIGURE 3.17 A comparison of oxygen incorporation levels of lightly doped p and heavily doped p and n (p^+ and n^+) silicon crystals grown under identical conditions. (From Oates, A. S. and Lin, W., *J. Cryst. Growth.*, 89, 117, 1988. Reproduced with permission from Elsevier.)

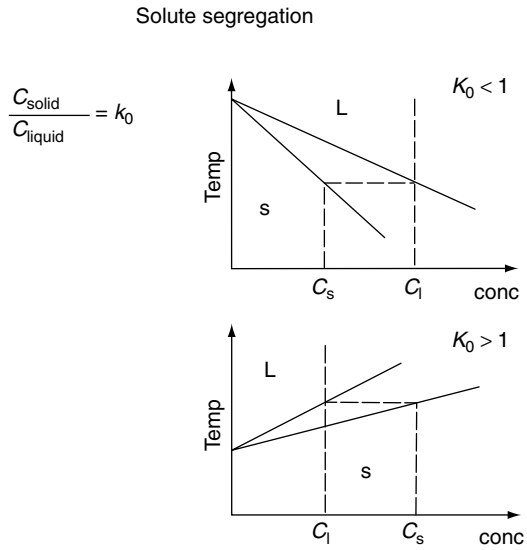


FIGURE 3.18 Schematic showing relationship between k_0 of solute and slopes of liquidus and solidus in a binary system. (From Lin, W., *Proceedings of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 1996, 288. Reproduced with permission.)

(thermal- and forced- convection) used and surface evaporation rate determine the melt oxygen concentration near the growing interface for incorporation [23]. Therefore, the oxygen in the silicon melt during CZ growth is a dynamic system. Depending on the growth parameters employed, the oxygen concentration and its profile along the grown crystal can be drastically different from those expected from a "normal freezing" behavior of the dopant element. Figure 3.16 shows oxygen concentration profiles from several CZ crystals grown using different parameters but with the same grower and setup. From these profiles, it appears that there is no common "segregation" behavior for oxygen displayed by these CZ silicon crystals. In fact, these axial oxygen distributions are not the results of oxygen segregation during silicon solidification. One can certainly fit the power function of Equation 3.1 to any of the profiles in Figure 3.16 and obtain a k value. But the k value so obtained does not describe the segregation behavior of oxygen during silicon freezing, but merely represent the characteristics of the oxygen concentration change in the melt, near the growing interface in a particular growing process. Therefore, it is inappropriate and incorrect to assign the k value so obtained as a "segregation coefficient" of oxygen. More importantly, it must be emphasized that the k value so obtained has absolutely no relationship with the equilibrium segregation coefficient of oxygen, which is a physical constant.

Although oxygen segregation is not visualized readily at the macro level, it is realized at the micro level at the growing interface in CZ silicon growth. In general, the degree of impurity incorporation (i.e., the effective segregation coefficient) at the growing interface is determined by the crystal growth rate f, at the interface, the boundary layer thickness δ (a function of crystal rotation) and equilibrium segregation coefficient k_0, as may be described by BPS expression, Equation 3.4. In the BPS expression, the k_0 and diffusion coefficient D are material constants, and f and δ are crystal growth parameters. If the crystal and crucible rotations are maintained constant, then the only parameter can cause variation in oxygen incorporation rate is the growth rate. An analysis of Equation 3.4 shows that the conditions listed in Table 3.1 are true. It is clear from the table that the impurity incorporation rate can fluctuate with the changing growth rate, unless the equilibrium segregation coefficient k_0 is unity. However, $k_0 = 1$ is not consistent with the phase rule [49]. Furthermore, it is also realized that the change in k_{eff} is greater when k_0 is further deviated from unity. In large diameter CZ silicon crystals, non-uniform oxygen

TABLE 3.1 Directions of Incorporation Rate Change in Response to Crystal Growth Rate Changes for $k_0 < 1$ and $k_0 > 1$

Equilibrium Segregation Coeff K_0	Crystal Growth Rate f	Impurity Incorporation Rate K_{eff}
>1	⇓	⇑
>1	⇑	⇓
<1	⇓	⇓
<1	⇑	⇑
$=1$		$k_{eff} = k_0$

There is no response in incorporation rate with growth rate change when $k_0 = 1$.

incorporation in the form of so-called oxygen striations are usually observed. The striations are results of impurity segregation due to fluctuations in the microscopic growth rate. The existence of oxygen striations in CZ silicon is an evident that oxygen in silicon does segregate and that k_0 for oxygen is not unity.

3.3.2.4.2.1 Equlibrium Segregation Coefficient, k_0, of Oxygen

The k_0 value of oxygen defines the basic segregation characteristics and therefore the effective incorporation rate during the silicon growth. The k_0 has been of interest to researchers in order to understand the microscopic oxygen incorporation and resulting precipitation characteristics in silicon. However, the details of the phase equilibrium of the Si–O system for the Si-rich alloys have not been extensively studied. Many investigations for the k_0 value of oxygen have only been carried out in the last three decades. The efforts can be divided into two categories. The first is a direct method, by oxygen concentration analysis on quenched oxygen–silicon alloys, as is commonly done in phase diagram studies. The second approach is to deduce the k_0 value from the silicon crystal growth experiments. One major factor affecting the k_0 determination is the accuracy of the oxygen concentration analysis. In this regard, infrared absorption using single crystal is more accurate and less uncertain than oxygen analyses on quenched samples using other methods such as differential thermal analysis. Due to the variety of methods used and accuracy of the oxygen concentration determination, the k_0 values reported over the years certainly do not show a great consensus. A range of k_0 values have been reported by various authors [27,50–55], ranging from greater to less than unity and including unity.

Among the various k_0 studies, the use of crystal growth experiments and analysis via the BPS relation has been considered a very accurate and repeatable approach. Lin and Hill [52] first applied the approach by growing a small diameter (14 cm) crystal in a large system in a controlled experiment. The oxygen incorporation level was observed to change with a change in crystal growth rate and it shows that oxygen does segregate during silicon solidification. From such an experiment, an equilibrium segregation coefficient of approximately 0.25 was deduced. Similarly, Lin and Stavola [53] studied the origin of microscopic oxygen inhomogeneity and its effect on oxygen precipitation using large-spacing oxygen striations prepared by a manually controlled crystal growth. The periodic nature of the oxygen profile shown in Figure 3.19 is a result of growth rate fluctuations and oxygen segregation. With $k_0 < 1$, the oxygen fluctuations and growth rate fluctuations are "in phase" (i.e., the maxima and minima of the oxygen fluctuations coincide with the microscopic growth rate fluctuations) see Table 3.1. The growth rate fluctuations are due to asymmetrical temperature distribution in a large silicon melt. These results again show that oxygen segregates significantly, which corresponds to a non-unity equilibrium segregation coefficient. The phase analysis [53] showed that $k_0 < 1$ for oxygen and that oxygen behaves similarly to arsenic in silicon (i.e., $k_0 \sim 0.3$). More recently, Iino et al. [55] used a similar approach to study the k_0 of oxygen by analyzing oxygen striations in comparison with that due to phosphorus. They obtained k_0 values between 0.13 and 0.37 (averaged to 0.21). The range of values obtained are in good agreement with the values obtained by Lin et al. [52,53].

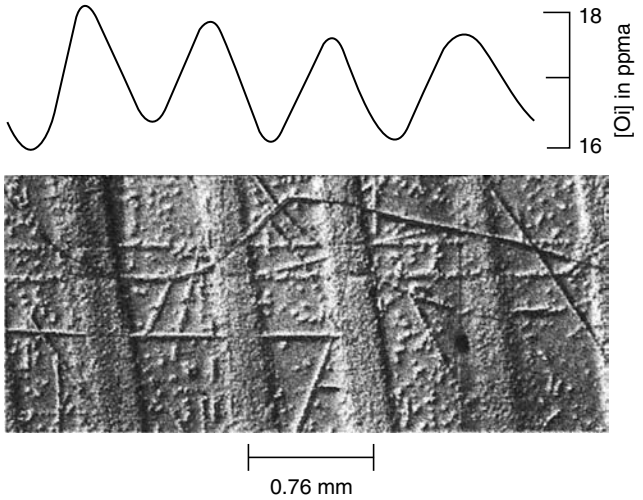


FIGURE 3.19 Top: IR 9 μm profile parallel to crystal axis showing periodic oxygen fluctuations. Bottom: micrograph of etched silicon after heat treatment showing precipitation bands corresponding to high oxygen regions of the fluctuations (top). (From Lin, W., *Proceedings of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 1996, 288. Reproduced with permission.)

3.3.2.4.2.2 Microscopic Inhomogeneity and Oxygen Precipitation

Oxygen segregation produces inhomogeneity when there is a perturbation in the growth rate. The thermal treatment experiments of silicon containing microfluctuations in oxygen concentration show that the precipitation is not uniform. Figure 3.19 bottom is a micrograph after the sample was heat treated at 1050°C for 5 h following chemical etching. It shows heavy precipitate bands corresponding to high concentration regions of oxygen in the fluctuations. The precipitation density in the low oxygen region appears to be insignificant. The difference in the precipitation densities in the high and the low oxygen regions, however, cannot be accounted for by the difference in their [Oi], approximately 9% (concentration fluctuation is ~4.5% about the mean). Furthermore, the fact that the precipitation occurs at the sample surface suggests that no denudation takes place. Apparently, the oxygen outdiffusion near the sample surface is suppressed by a fast nucleation/growth mechanism in the high oxygen band regions, while the precipitation kinetics at the low oxygen regions seem retarded. These preferential precipitation bands correspond to concentric ring patterns often observed in the oxidized- or heat-treated wafers following chemical etching. An example is shown in the x-ray topography in Figure 3.20.

The oxygen precipitation kinetics in CZ silicon with microfluctuations is not strictly proportional to oxygen concentration. It is likely that the preferential precipitation of oxygen in the high [Oi] regions reflects the large number of nuclei available in the same regions. It is reasonable to postulate that the mechanism for inhomogeneous precipitation behavior involves the preferred nucleation of the precipitates at microdefect centers introduced during crystal growth. Such defects, ranging from several hundred angstroms to tenths of a micron in size, have been related to temperature fluctuations and remelting phenomenon at the growing interface.

The grown-in microdefects in CZ silicon have been extensively studied in recent years. Basically, as in the FZ silicon, two types of microdefects may be formed during crystal growth [58]; D defects are vacancy agglomerates in nature and their formation and density are growth rate dependent. The D defects form when the growth rate is above a critical value, below which the A-defects formation are favored. The A defects are clusters of silicon self-interstitials in the form of dislocation loops. In a crystal's "D defect region," where the vacancy concentration is excessive (greater than the equilibrium value), defects such as

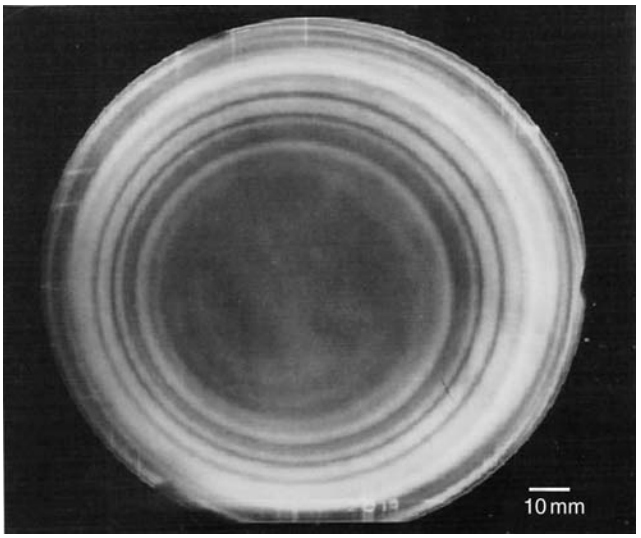


FIGURE 3.20 X-ray topography of heat treated CZ wafer showing concentric oxygen precipitation bands at high oxygen regions of the concentration fluctuations. (From Shimura, F., *Semiconducto Silicon Technology*, Academic Press, New York, 1989, 258. Reproduced with permission from Elsevier.)

crystal originated pits (COP), flow pattern defects (FPD) and laser scattering tomography defects (LST) have been observed and they have been shown to cause degradation of MOS devices (mainly gate oxide integrity) [59]. Further discussion on grown-in microdefects and their relevance to crystal growth parameters will be made later in this chapter. Nakajima et al. [60] studied the distribution of as-grown D defects in the silicon crystal in relation to microscopic growth rate fluctuations on large striated silicon crystals. The study revealed that LST defects occur at the maxima of the growth rate fluctuations, while FPD defects occur at the minima. See Figure 3.21. This result means that the LST defects occur at the high oxygen concentration regions of the $[O_i]$ fluctuations. The LST defects have been speculated as oxygen precipitates in nature (they can be annihilated by high-temperature anneal in hydrogen) [61]. It is assumed that its formation may be the result of either (a) nucleation of oxygen precipitates on vacancy clusters or (b) direct interaction of oxygen atoms with non-clustered point defects. The existence of LST defect centers (a current crystal as shown in Figure 3.19 was grown with a growth rate of 1.5 mm/min) suggests there are far more nuclei available in the high oxygen regions than in the low oxygen regions. When subjected to precipitation heat treatment(s), as the oxygen precipitation progresses in the bands of high nuclei density, self-interstitials are ejected and flood the neighboring low oxygen regions, where the precipitation has not started. The supersaturated self-interstitials would raise the nucleation barrier in the low oxygen regions of the fluctuations and retards its precipitation process [62].

In the current discussion, the occurrence of the microdefects in the high oxygen regions may be a major factor in the formation of observed preferential precipitation. It would be interesting and useful to conduct a similar experiment with a crystal grown with microfluctuations in oxygen, but without D defects by using a low-growth rate. Such an experiment would clarify the role of microdefects in the observed preferential oxygen precipitation.

3.3.2.4.3 Controlled Oxygen Silicon Crystal Growth

3.3.2.4.3.1 Normal Czochralski Growth

For ease of discussion, oxygen concentrations in silicon crystals for IC applications may be conveniently classified into high, medium, and low concentration ranges. If we designate 14–17 ppma (ASTM -F121-80)

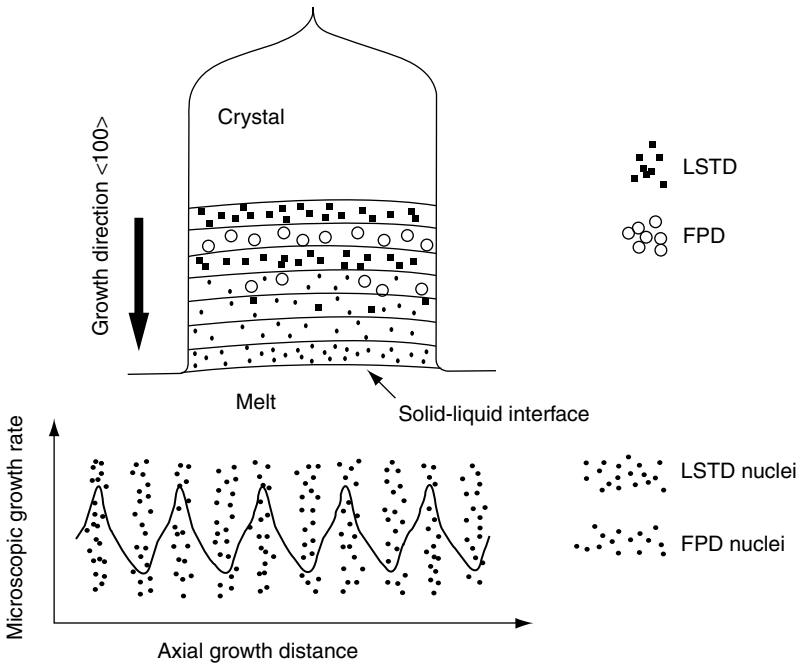


FIGURE 3.21 Schematic showing the occurrence of laser scattering tomography (LST) and flow pattern density (FPD) defects at maxima and minima of the growth rate fluctuations, respectively, of the CZ growth. (From Nakajima, K. et al., *Semiconductor Silicon 1994*, eds. Huff, H. R., Bergholz, W., and Sumino, K., Electrochemical Society, Pennington, NJ, 1994, 168. Reproduced with permission from Electrochemical Society.)

as the medium range, the concentrations above and below this range are referred to as high and low concentrations, respectively.

From previous discussions, it is seen that the forced convection is an effective tool for controlling oxygen incorporation. In order to achieve a desired oxygen level with an axial uniformity in a silicon crystal, the following procedure may be carried out. Experimentally, for a given crystal growing system, one can establish oxygen incorporation profiles as a function of crystal/crucible rotation rates via studies, such as shown in Figure 3.15. Using selected rotational parameters at different stages of crystal growth, one can develop and tailor the growth processes to grow crystals of desired oxygen concentration with substantial axial- and radial-uniformity. Figure 3.22 shows an example of using variable crucible rotation rates to change the oxygen incorporation levels during the growth, while the crystal rotation rate is maintained constant. It shows that the crucible ramping to a higher rotation rate effectively raises the oxygen level and changes the oxygen concentration profile. The incorporation level can be further enhanced when alternate ramping-up and -down of crucible rotation at medium and high rates are employed, as shown in the Figure 3.22. Presumably, the action causes local "disturbance" and thinning of the boundary layer between the crucible and the melt, thus increases crucible dissolution. The example in Figure 3.22 demonstrates the usefulness of forced convection in the enhancement or retardation of oxygen incorporation. With the proper use of forced convection, including the use of alternate ramping-up and -down of the crucible rotations, the increased uniform axial incorporation for high and medium ranges of oxygen can be achieved. Concentration profiles a and b of Figure 3.23 are oxygen profiles using variable crucible rotations for high and medium oxygen concentration levels.

In the crystal growing systems, where the forced convection alone cannot achieve the level/uniformity desired, additional sources of oxygen may be added by increasing the melt-crucible contact surface.

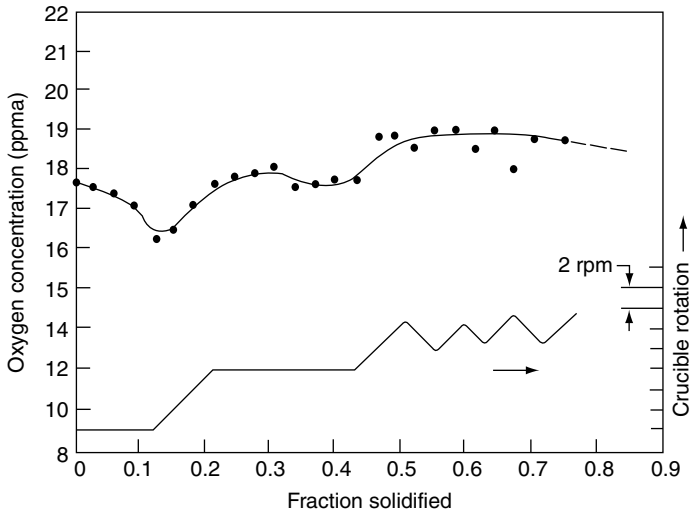


FIGURE 3.22 Axial oxygen profile of a silicon crystal showing the effect of crucible rotation rate on oxygen incorporation level. (From Lin, W. and Benson, K. E., *Annual Review of Materials Science*, 17, 273, 1987. Reproduced with permission from Annual Reviews.)

Methods such as sand-blasting of the crucible surface [63] and the addition of an extra quartz rod/ring placed at strategic locations [64] have been used. Figure 3.24 shows a crucible design with its bottom surface fabricated in a corrugated configuration. It is shown that the additional contact sources uniformly increase the incorporation level, curve b of Figure 3.25, as may be compared with curve c,

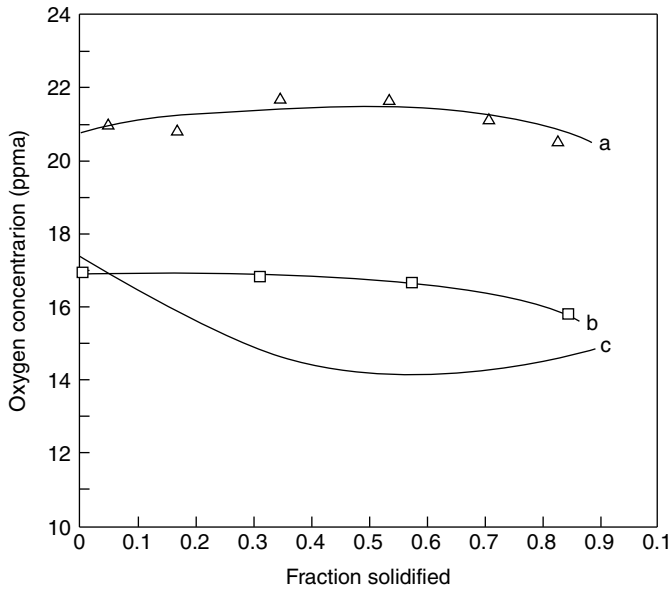


FIGURE 3.23 Uniform axial oxygen profiles of silicon crystals using variable crucible rotation rates during growth (curve a and b), as compared with that due to normal growth (curve c). (From Lin, W., *Oxygen in Silicon*, ed. Shimura, F., Academic Press, 1994, chap. 2. Reproduced with permission from Elsevier.)

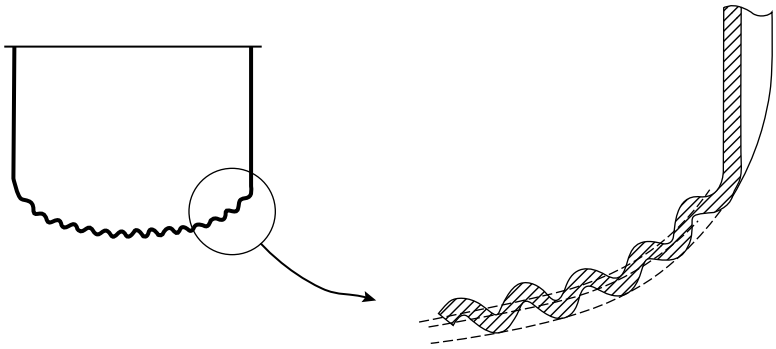


FIGURE 3.24 Schematic of the cross-sectional view of the crucible design with corrugated crucible bottom. (From Lin, W., *Oxygen in Silicon*, ed. Shimura, F., Academic Press, 1994, chap. 2. Reproduced with permission from Elsevier.)

grown with a normal crucible. Curve a shows an oxygen profile resulting from the silicon growth with extra quartz material adhered to the crucible bottom surface as an added oxygen source under normal growth conditions. It is pointed out that profiles a, b, and c are obtained for a fixed set of crystal and crucible rotations. When variable forced convection parameters are applied, the concentration profiles may be tailored to improve axial uniformity as shown in Figure 3.23. Curve a of Figure 3.23 is the result of using variable crucible rotation on crystal growth with additional quartz material adhered to the crucible bottom.

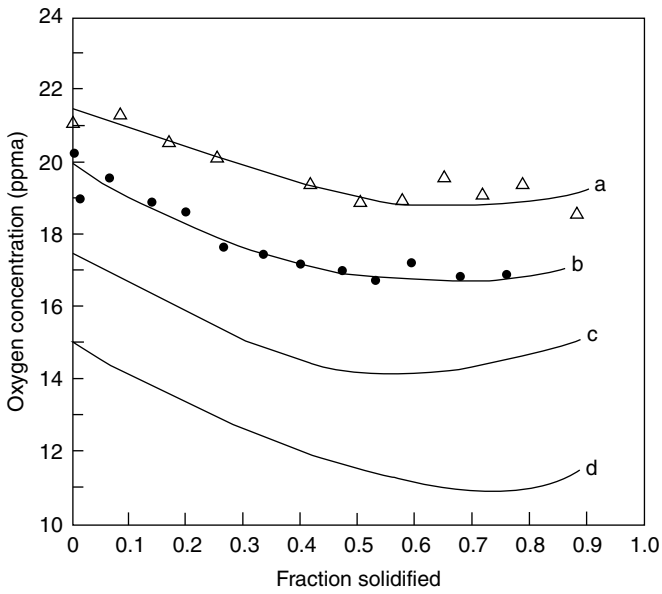


FIGURE 3.25 Axial oxygen profiles showing enhanced incorporation by (a) extra quartz adhered to crucible bottom and (b) corrugated crucible bottom. Curve (c) is due to normal growth and curve (d) is due to the use of forced convection conditions for low-oxygen incorporation. (From Lin, W., *Oxygen in Silicon*, ed. Shimura, F., Academic Press, 1994, chap. 2, Reproduced with permission from Elsevier.)

While thermal convection, crucible, and crystal rotations all have effects on the oxygen incorporation level, these parameters also have a great impact on the radial oxygen gradient. The relevance of these parameters to radial impurity uniformity has been discussed in Section 3.3.2 of this chapter. In order to achieve radial oxygen uniformity, the melt flows causing the radial non-uniformity in diffusion boundary layer thickness have to be suppressed. For example, when using a high crucible rotation rate to enhance oxygen incorporation, the radial oxygen uniformity is degraded unless, a high crystal rotation in the opposite direction is used. As discussed earlier, the crucible rotation creates a fluid flow in the same general direction as the thermal convection (see Figure 3.7c). Such flow will cause the diffusion boundary layer to be thinner at the edge than at the center of the interface. Consequently, more oxygen is incorporated at the center than at the edge of the crystal. The increased crystal rotation suppresses the effect of crucible rotation and thermal convection flows, and results in a thinner diffusion boundary layer. The thickness variations in the thinner boundary layer would cause less radial gradient in oxygen incorporation. The faster crystal rotation also results in more mixing in the melt, adjacent to the growing interface and tends to improve the melt concentration uniformity. It should be mentioned that there exists a drastic decrease in oxygen concentration in the CZ crystal's radial oxygen profiles near the crystal's periphery region. This is due to oxygen out-diffusion during the crystal growth following solidification.

In large melt growth, the controllability for low level oxygen incorporation is more limited than that for the medium and high incorporation levels, due to the domination of thermal convection function as the transport of oxygen-rich melt. This is enhanced for the growth of the crystal portion near the seed end. The curve d in Figure 3.25, represents a typical range of concentration profiles in the low oxygen range (11–14 pima) that may be obtained with the experimental hot-zone used. In this range, further uniformity of the profile can be obtained by increasing the incorporation level of the lower concentration portion of the profile using crucible ramping. However, suppressing the seed-end oxygen incorporation by forced convection is usually not an efficient process. It often results in degradation in the oxygen radial gradient and other disadvantages in growth conditions and crystal properties.

From the above discussion, one realizes that in normal CZ growth, the small and/low-aspect-ratio melt configurations facilitates the incorporation of low-levels of oxygen due to the reduction of thermal convection. Such a growth environment may be obtained by growing silicon from a small/shallow inner crucible in a double-crucible type set-up. Axial uniformity as well as concentration control using a double-crucible set-up has been demonstrated [65]. Although, such an apparatus is more involved than a normal CZ growth, as the CZ charge size continues to increase, the concept of a small/shallow melt growth in a double-crucible is capable of extension to continuous CZ growth, to be discussed later in this section. For silicon materials requiring low oxygen and low microdefect density, however, the most effective method is growth under an applied magnetic field. For very low level oxygen incorporation (a few parts per million atomic), crucibles made of non-oxygen containing materials, such as Si_3N_4, have been used for silicon crystal growth [66,67]. In this case, the source of oxygen is from the silica beads added in the silicon melt. The oxygen incorporation is controlled by the ratio of the surface area of the SiO_2 beads to the free melt surface. In general, the oxygen incorporation in normal CZ silicon growth from a large melt can be controlled within ± 1.5 ppma of the normal concentration range of interest.

3.3.2.4.3.2 Magnetic Field Applied Czochralski Growth (MCZ)

From the previous discussion, it is clear that the thermal convection in a large silicon melt plays a major role in determining many aspects of the crystal quality of CZ silicon. In particular, oxygen concentration as well as dopant and oxygen uniformity are of concern. The ability of a magnetic field to suppress the thermal convection in electrically conducting fluids was demonstrated in the 1960s in crystal growth experiments [68]. In 1970, a transverse magnetic field was applied for the same purpose in CZ growth of indium antimonide [69]. The application of a magnetic field across an electrically conductive melt increases the effective kinematic viscosity of the melt [70] and thus, suppresses the thermal convection and related temperature fluctuations in the melt.

Hoshi et al. [71] first reported CZ silicon crystal growth under an applied transverse magnetic field. Aside from a reduction/elimination of impurity striations, MCZ displayed the potential for growing low-oxygen, low-microdefect, and higher resistivity CZ silicon for applications in power devices and imaging devices such as CCD. Since 1980, various types of magnetic field configurations [72,73] in terms of field directions (VMCZ for vertical, HMCZ for horizontal, and cusp magnetic field) and the magnet types used (normal conductive or superconductive), have been developed and crystal growth studies have been carried out. Figure 3.26 shows a comparison of the magnetic field direction/distribution of the three MCZ methods.

In general, both VMCZ growth and HMCZ growth have been shown to reduce temperature fluctuation and related growth rate fluctuations, resulting in reduced impurity striations. However, the magnetic field effects on impurity incorporation behavior vary widely depending upon the field direction with respect to the growth axis and the field strength. For example, increased field strength enhances oxygen concentration under a vertical field [74], whereas oxygen incorporation is retarded under an increased transverse magnetic field [75]. Therefore, a transverse magnetic field facilitates the

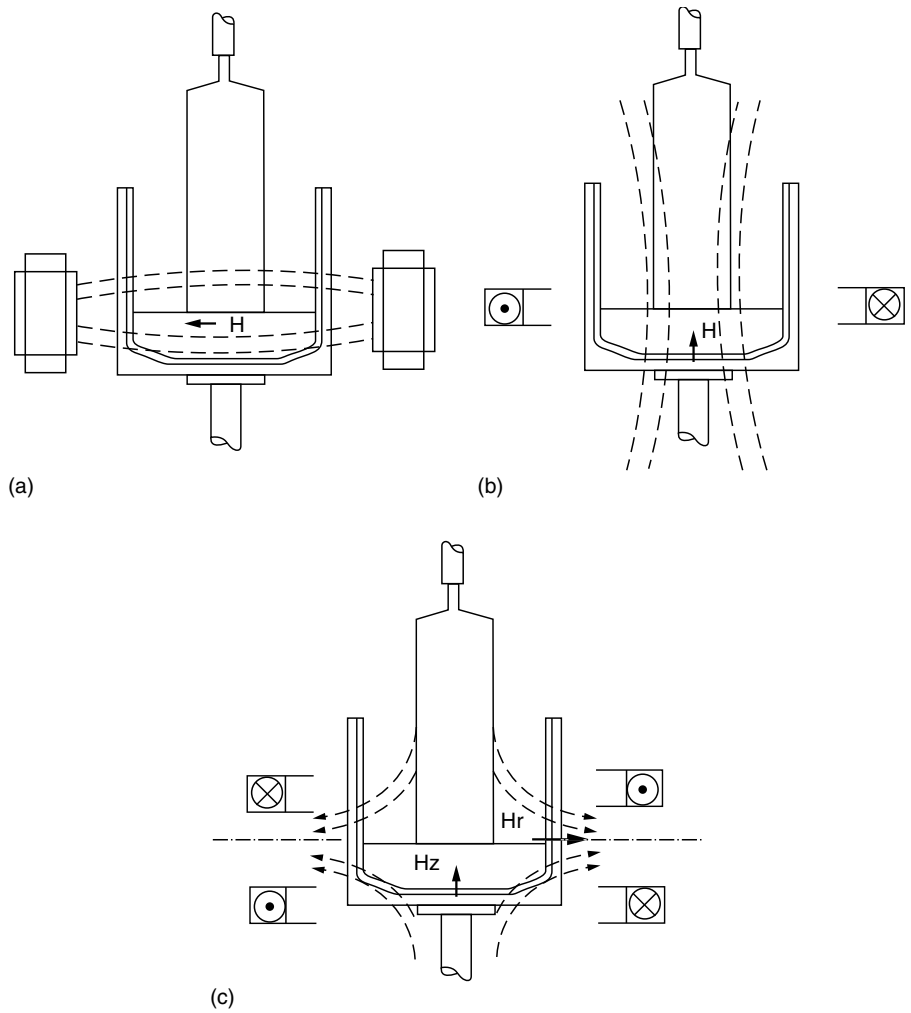


FIGURE 3.26 Schematics showing arrangements for magnetic Czochralski; (a) horizontal magnetic field, (b) vertical magnetic field and (c) cusp magnetic field. The arrows indicate field directions.

incorporation of low ($<5\times10^{17}$ atoms/cm^3) to medium ($5–10\times10^{17}$ atoms/cm^3) oxygen concentrations, whereas a vertical field facilitates the growth of silicon with medium to high ($>10^{18}$ atoms/cm^3) oxygen concentrations. In fact, extremely high oxygen concentrations, near or above the solubility limit, can be incorporated under certain growth conditions with vertical magnetic fields [76,74]. Forced convection induced by crystal and crucible rotations can also perturb the oxygen incorporation under a magnetic field. Several large diameter MCZ crystal growth experiments from large melts have provided information on the difference in oxygen incorporation behavior between HMCZ and VMCZ. For example, Ohwa et al. [76] made a comparison of oxygen incorporation behavior between VMCZ and HMCZ using the same puller, as is shown in Figure 3.27. In the HMCZ mode, a low level of oxygen concentration (~4 ppma) is incorporated in a horizontal magnetic field of 0.25 Tesla (2500 G) with good axial uniformity. The crystal rotation shows no effect on the incorporation level under this condition. However, it is shown in another study [75] that, under HMCZ, an increase in crucible rotation increases the oxygen incorporation level. As in the normal CZ, high crucible rotation often results in some degradation in radial uniformity. In such an instance, the use of a higher crystal rotation rate would help to reduce the oxygen radial gradient. Figure 3.27 also shows that, in VMCZ mode, higher oxygen concentration may be incorporated using a magnetic field strength one-third of that for HMCZ. Unlike HMCZ, the oxygen incorporation level in VMCZ is a strong function of crystal rotation and the

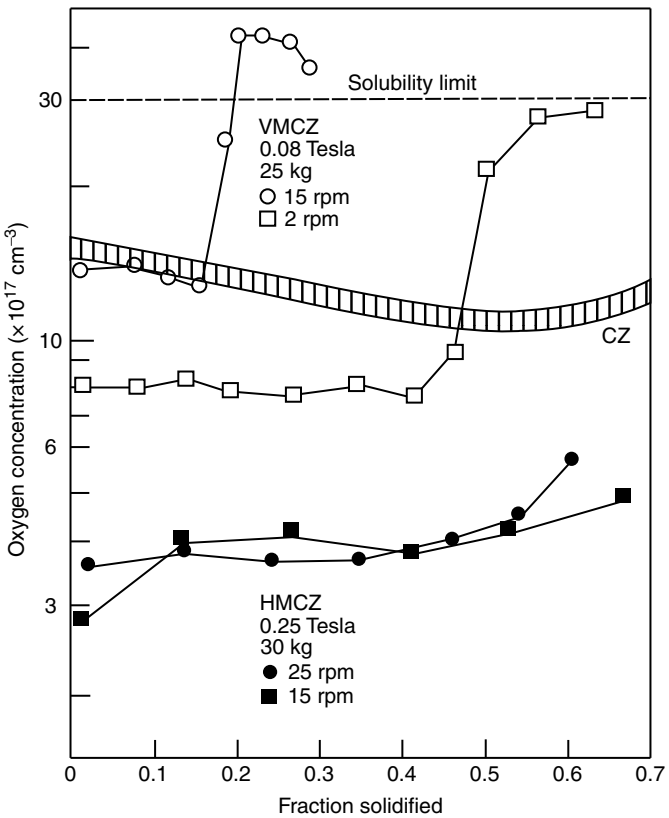


FIGURE 3.27 Oxygen profiles of silicon crystals grown under horizontal and vertical applied magnetic fields under the conditions indicated. The data show that crucible rotation rate has significant enhancement effect on oxygen incorporation under vertical magnetic field. (From Ohwa, M. et al., *Semiconductor Silicon 1986*, eds. Huff, H. R., Abe T., and Kolbesen, B., Electrochemical Society, Pennington, NJ, 1986, 117. Reproduced with permission from Electrochemical Society.)

incorporation switches to a much higher level through a sharp transition during the crystal growth. Both the incorporation level and the transition point are clearly a function of crystal rotation, indicating that under VMCZ mode, the crystal rotation induced forced convection acts as a major transport mechanism.

Figure 3.28 shows models proposed by Hoshi et al. [77] for the magnetic damping of the melt flow for VMCZ and HMCZ. In view of the model for VMCZ, the melt flow from outer to center regions of the melt (flow perpendicular to the vertical magnetic flux) are retarded. Under this condition, the forced convection induced by the crystal rotation is an effective transport for the oxygen-rich melt from the crucible bottom to the growing crystal. As the melt aspect ratio is reduced during the crystal growth, at some transition point, the forced convection would take over as the dominant transport mechanism and the incorporation level is sharply increased. The occurrence of this "transition" seems to depend on the crystal rotation used, the higher the crystal rotation, the sooner this transition occurs during the growth. It is interesting to note that the observed "transition" in incorporation level for the VMCZ growth is similar to that discussed earlier for normal CZ growth under high crystal rotation (30 rpm) with the crucible in iso-rotation mode (2 rpm) (see Figure 3.14). In both cases, the behavior is attributed to strong crystal-rotation-induced convection. In the case of HMCZ, the model of Hoshi et al. shows that the transverse magnetic flux damps the vertical melt flow near the wall, due to thermal convection and crucible rotation, resulting in retarded oxygen transport and therefore, a low level of oxygen incorporation. The forced convection induced by the crystal and crucible rotations is more effective in the transverse direction, parallel to the magnetic flux. Thus, the increased crucible rotation would help crucible dissolution and transport of the oxygen-rich melt follow flow path delineated in the model.

The results by Ohwa et al. discussed above and other large-diameter, large-melt VMCZ growth indicate that oxygen incorporation (and other growth characteristics) in VMCZ is affected by many variables and is not easily controlled. Furthermore, impurity segregations were observed to depend on the vertical field strength [78,79]. The segregation coefficient of the impurity (such as phosphorus, gallium, carbon, and oxygen) tends to increase (towards unity) as the magnetic field increases. However, the radial uniformity of these impurities is significantly degraded in a vertical field [74]. For example, the radial oxygen gradient in a 100-mm-diameter silicon crystal can reach 30%–50% when grown in a high vertical field. These properties are not desirable. The change in segregation and severe degradation of the radial

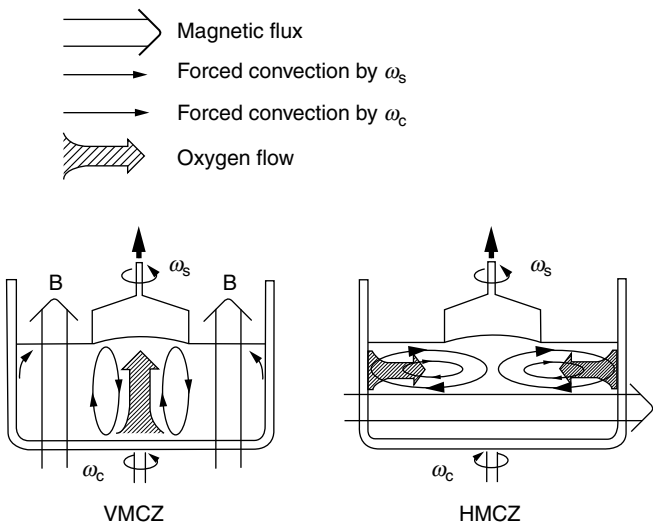


FIGURE 3.28 The effect of vertical and horizontal magnetic field on damping melt flow in Czochralski crucible. (From Shimura, F., *Semiconducto Silicon Technology*, Academic Press, New York, 1989, 258. Reproduced with permission from Elsevier.)

gradient found for growth in VMCZ has not been reported for HMCZ growth. From considerations of growth control ability and crystal properties, HMCZ is the preferred approach for MCZ growth. The HMCZ method can grow large diameter silicon with oxygen levels ranging from a few parts per million atomic to over 20 ppma with axial concentration uniformity.

The CZ crystal growth under an applied cusp magnetic field was designed to minimize the undesirable characteristics of VMCZ discussed above [80,81]. A CZ growing system with cusp magnetic field uses two sets of coils (often superconducting) co-axially with the crystal, which are energized in opposing directions [82], as shown in Figure 3.26c. In this arrangement, the crystal-melt interface is located in the symmetry plane between the two coils and is maintained in this position throughout the growth by adjusting the crucible height. The resulting magnetic field distribution in the growing system is represented by the dotted lines in Figure 3.26c. Essentially, the major and significant magnetic components are vertical magnetic field orthogonal to the crucible bottom, H_z, and the radial field orthogonal to the sidewalls of the crucible, H_r. The melt free surface has no orthogonal magnetic field component. As discussed earlier, the crucible erosion rate and evaporation rate at the melt free surface are the factors determining the oxygen concentration in the bulk melt, and therefore the incorporation level. The erosion rate at the sidewall is believed to be under diffusion control. The orthogonal components of the magnetic field at the sidewall and bottom have the damping effect on melt flows parallel to the crucible surface (such as thermal convection flows), resulting in thicker diffusion boundary layer, and thus a lower erosion rate. At the melt free surface, on the other hand, the boundary layer controlling the oxygen evaporation is determined by the radial melt flows, such as Marangoni flows and centrifugal flows pumped outwards from the crystal by the crystal rotation induced forced convection. But these flows are not damped since there are no orthogonal magnetic field components at the melt free surface. Thus, the oxygen level can be reduced by retarding the crucible erosion while leaving surface evaporation unaffected. The degree of oxygen level reduction is proportional to the applied field strength generated by the two coils. However, when the melt surface is located away from the symmetry plane of the cusp magnetic field, the melt surface will be subjected to the effect of orthogonal magnetic components, resulting in reduced oxygen evaporation and increased oxygen level [79].

In general, the main application of cusp magnetic field applied CZ growth is in the low oxygen incorporation in addition to the usual advantages of magnetic CZ, such as reduced temperature fluctuations in the melt, etc. The orthogonal components acting on the melt surface and on the crucible walls can be controlled independently. Thus, the controllability of the magnetic field is the main benefit such that the desired incorporation level with good axial uniformity can be achieved. The crystals so grown are without degradation in radial properties and abnormal segregation behavior of the dopant experienced by the VMCZ incorporation. However, it is essential to optimize the crystal and crucible rotation condition to match the applied field strength and fraction of solidification, during the growth, in order to achieve radial and axial uniformity of oxygen and dopant [83].

3.3.3 Grown-In Microdefects

3.3.3.1 Relevance to Growth Parameters

Microdefects are formed in normal "dislocation-free" melt-grown silicon crystals due to condensation or agglomeration of excess point defects at or near the growing interface. In the 1970s, the so-called A and B defects were observed in FZ crystals [84]. The A defects were identified as small "extrinsic" dislocation loops and B defects the embyros for the A defects as characterized by transmission electron microscopy (TEM) [164]. The defects are the result of condensation of Si self-interstitials and they can be electrically active when decorated, as may be revealed by EBIC (electron beam-induced current) [85]. Fast growth rates were found to eliminate the A and B defects [84]. The other type of defects, termed "D" defects in FZ crystals were attributed to the condensations of excess vacancies. In recent years, the microdefects in as-grown CZ silicon have been studied extensively. Defects similar to D and A defects found in FZ silicon have been observed in CZ silicon [86]. The major findings and their relevance to crystal growth are discussed in the following.

In a CZ silicon grown with a "normal" growth rate, the cross-section can often be divided into two regions: the inner region is vacancy-rich and the outer concentric region is silicon self-interstitial rich. The two regions are separated by a concentric ring border that can be characterized by populations of oxidation induced stacking faults (OSF) upon oxidation of the wafer as is shown in the schematic in Figure 3.29. Empirically, it has been found that the diameter of the "OSF ring" is reduced when the growth rate is reduced and vice versa. The microdefects have been observed inside the OSF-ring by several detection methods; Crystal originated pits (COPs) as revealed by repeated standard clean 1 (SC1) of the wafer surface (as small scattering centers) [87] (see Figure 3.30), flow pattern defects (FDPs) as delineated by the Secco etch of the wafer surface without agitation [88], laser scattering tomography (LST) defects [89] or by an optical precipitate profiler (OPP). Their respective defect densities have been correlated with the crystal growth rate (Figure 3.31) and with the defect densities of the gate-oxide integrity (GOI) measurements (Figure 3.32). Furthermore, from the correlation such as shown in Figure 3.33, it was speculated that the defects detected inside the OSF-ring by different methods have a common origin, i.e., they are agglomerates of vacancies in nature and are referred to as "D" defects. Systematic studies have found that COPs, FPDs, and LSTs are same defects revealed by different detection methods [90]. The defects in the region outside the OSF-ring are referred to as "A" defects, which have been shown not to affect GOI, but degrade DRAM performance and yield. "A" defects are normally observed as "large pits" following Secco etch of the wafer surface.

Due to the harmful effect of the D defects on the gate oxide quality and A defects on DRAM, it is desirable to minimize the formation of clusters of either of these point defects, but to achieve a nominal low level of each type. In the extreme case, the diameter of the OSF-ring shrinks to zero and whole wafer is free of "D" defect. Experimentally, the OSF-ring diameter was found to increase as the growth rate is increased. However, the dependence of the ring diameter on the pull rate behave remarkably different for different crystal diameters or when the surface cooling condition is changed via, for example, an added radiation shield [91]. A radiation shield serves to isolates the crystal surface from the radiation heat from the heater and silicon-melt surface, resulting in an increase in surface cooling rate and therefore, an

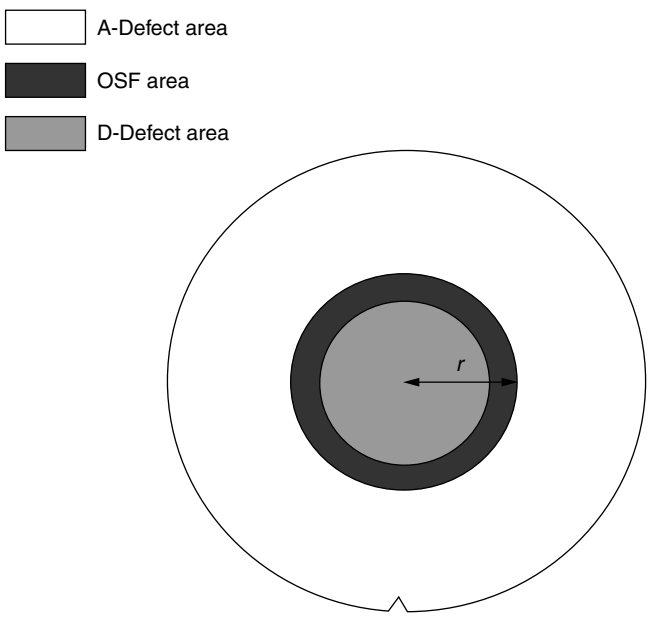


FIGURE 3.29 Schematic showing a ring region (oxidation induced stacking faults region upon oxidation) separating A-defect-rich and D-defect-rich areas.

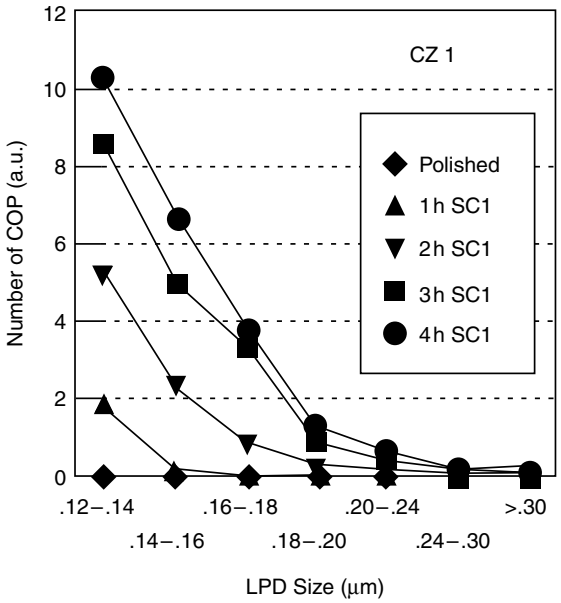


FIGURE 3.30 Light scattering point counts due to crystal originated pits (COPs) defects as a function of length of standard clean 1 (SC1) clean solution treatment and "particle size." (From Wagner, P. et al., *Proceeding of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 1996, 101. Reproduced with permission.)

increase in axial thermal gradient in the crystal near the liquid–solid interface. The net effect depends on the heat shield design. Crystal growth experiments with various designs of heat shields show that the OSF-ring diameter, and therefore the fraction of the silicon wafer containing D defects, depends not only on the crystal growth rate, but also the axial thermal gradient in the crystal at the growing interface. Based on the experiments, von Ammon et al. [91] found that there exists a constant critical value of $V/G(r) =$

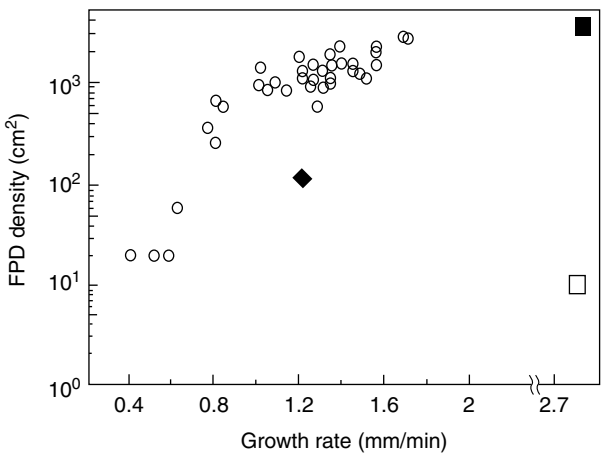


FIGURE 3.31 Flow pattern defect density as a function of CZ crystal growth rate. (From Yamagishi, H. et al., *Semiconductor Silicon 1994*, eds. Huff, H. R. and Bergholz, W., and Sumino, K., Electrochemical Society, Pennington, NJ, 1994, 124. 58. Reproduced with permission from Electrochemical Society.)

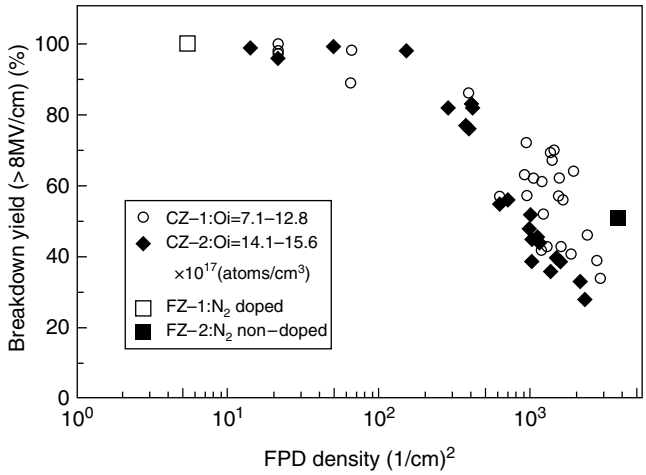


FIGURE 3.32 C-mode oxide break yield as a function of flow pattern defect density. (From Yamagishi, H. et al., *Semiconductor Silicon 1994*, eds. Huff, H. R., Bergholz, W., and Sumino, K., Electrochemical Society, Pennington, NJ, 1994, 124. Reproduced with permission from Electrochemical Society.)

$C_{crit} = 1.3 \times 10^{-3}$ cm^2 min^{-1} K^{-1}, which allows a calculation of the radial position r of the OSF-ring, if the pull rate V and radial variation of the axial temperature gradient $G(r)$ are known [91,92]. The empirical formula can be explained based on Voronkov's theory [93,94] which predicted a change from vacancy to Si interstitial type defects at a critical value of V/G. Based on the empirical equation, Figure 3.34 shows a computed dependence of OSF ring diameter on the crystal pull rate for various crystal diameters. It shows that the larger the crystal diameter, the slower the pull rate is required to completely eliminate the vacancy type defects [95]. At 100-mm-diameter, a pull rate of 0.65 mm/min is required for vacancy defect-free growth, while a much lower pull rate, 0.35 mm/min, is required to

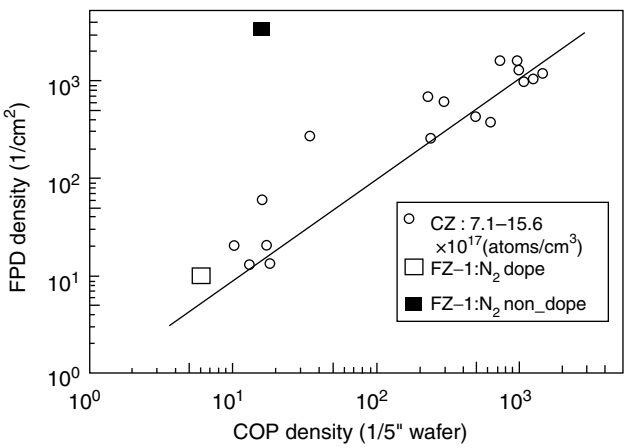


FIGURE 3.33 Correlation between COPs and FPDs. (From Yamagishi, H. et al., *Semiconductor Silicon 1994*, ed. Huff, H. R., Bergholz, W., and Sumino, K., Electrochemical Society, Pennington, NJ, 1994, 124. Reproduced with permission from Electrochemical Society.)

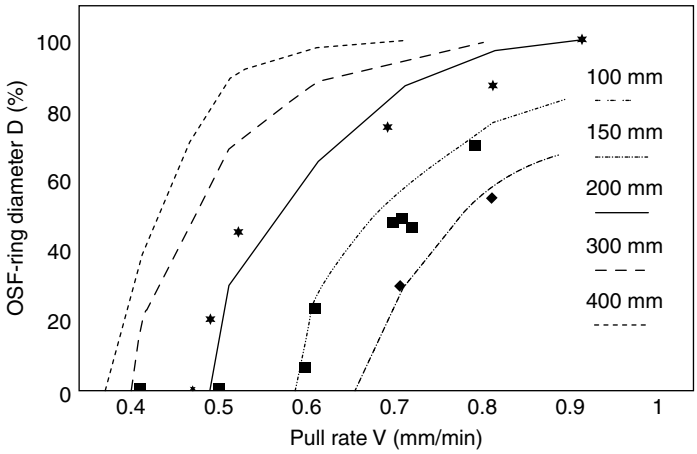


FIGURE 3.34 Oxidation induced stacking faults-ring diameter as a function of crystal pull rate for different crystal diameters. (From von Ammon, W., *Proceedings of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 1996, 233. Reproduced with permission.)

eliminate the vacancy defects for the 300-mm-diameter crystal. Too low a pull rate, however, presents a problem in maintaining a stable and steady growth rate, which is essential for dislocation-free growth.

The above discussions on the occurrence of the D defects are for lightly doped silicon crystal growth. The defect formation mechanism is apparently different for heavily doped boron silicon. It is found that the diameter of OSF ring [95] and COP density [96,97] decrease dramatically with increasing boron concentration (for resistivity <20 m ohm–cm, Figure 3.35). This effect was attributed to Fermi-level effects of boron-vacancy pairs [98]. The other probable explanation was based on the modification of the generation mechanism and/or diffusivity of the intrinsic defects in the strained silicon lattice, induced by the high concentration of smaller-sized boron atoms compared to silicon.

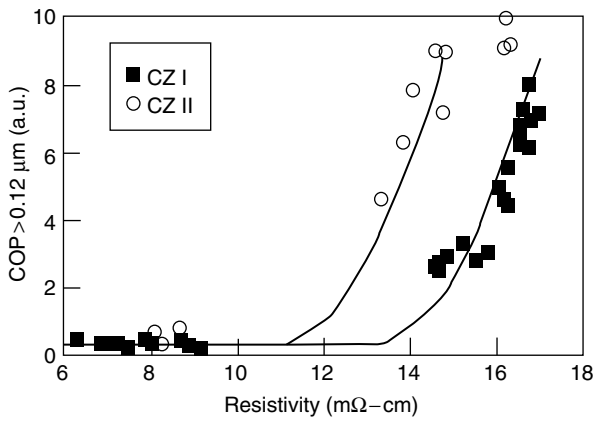


FIGURE 3.35 Crystal originated pits density as a function of crystal resistivity, showing that the crystal is essentially COP-free when the resistivity is less than 10 m ohm–cm. (From Wagner, P. et al., *Proceedings of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 1996, 101. Reproduced with permission.)

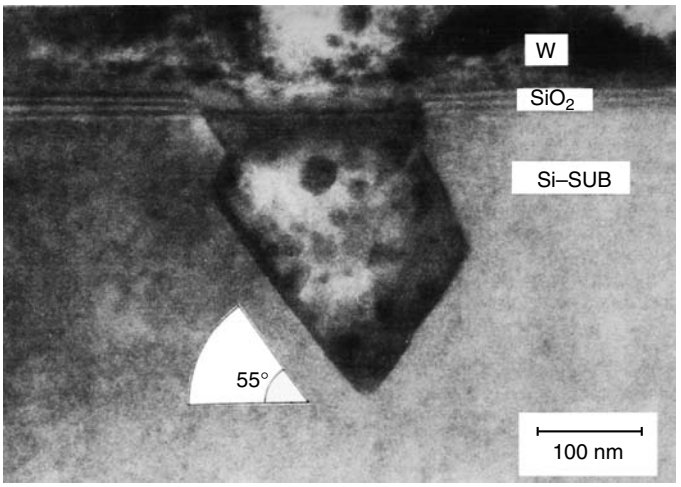


FIGURE 3.36 Cross-sectional transmission electron microscopy (TEM) observation of an oxide defect. (From Itsumi, M. et al., *Proceeding of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 1996, 270. Reproduced with permission.)

3.3.3.2 Defect Structures

Although the microdefects in CZ silicon have been detected for quite sometime by various methods as mentioned above, only recently the detailed structures have been revealed by TEM and it is meaningfully correlated to the features detected by the other methods. Modern day bulk defect detection methods, such as IR Laser Scattering Tomography and precision thinning/etching tools, such as FIB (Focused Ion Beam) make it possible to isolate bulk defects in a small thin sample to be studied by TEM. Since the most significant impact of the microdefects in silicon is on the GOI failures, the defects were first observed by TEM as voids by Parks et al. [99], as the origin of oxide defects. The octahedral voids were also observed to exist under the dielectric breakdown sites of thermal oxide films by the Cu decoration method [100], Figure 3.36. The TEM also identified the LST defects in the silicon

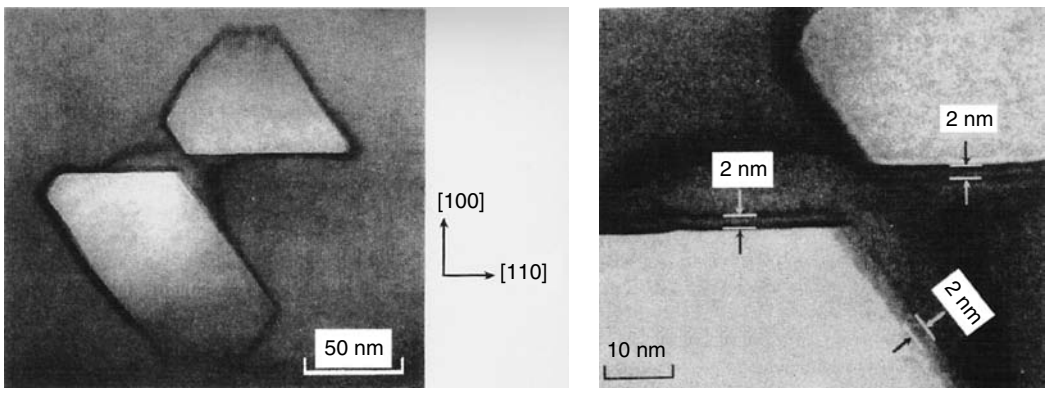


FIGURE 3.37 Cross-sectional TEM view of LST defects Photo in (a) is an enlarged view of (b) showing oxide layer at void-crystal interface. (From Itsumi, M. et al., *Proceedings of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 1996, 270. Reproduced with permission.)

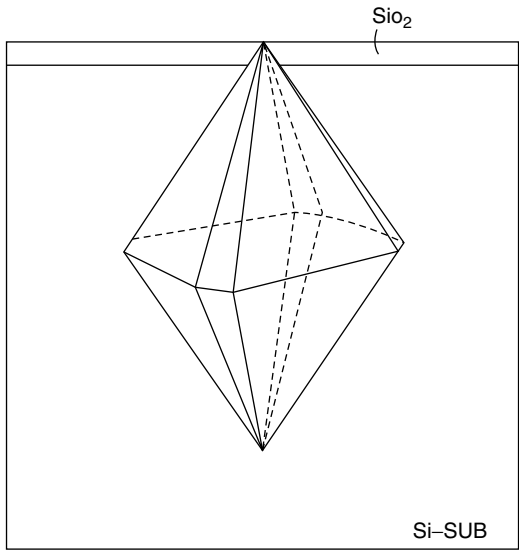


FIGURE 3.38 Schematic representation of an octahedral defect structure. (From Itsumi, M. et al., *Proceedings of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 1996, 270. Reproduced with permission.)

bulk as octahedral voids. [101–103]. In general, the TEM findings on the D defects can be summarized as voids bounded by {111} planes having a structure of two inverted pyramids, Figure 3.37 and Figure 3.38. The analyses by the energy dispersive x-ray spectroscopy (EDXS) indicated that there are no other signals except silicon. The silicon signal is weaker inside the defect than its surrounding, an evidence of a cavity. Their typical size is 0.1–0.2 μm and they often appear as twins or triplets, although the individual defect is usually an incomplete octahedron (some tops of the octahedrons are cut to result in a complex polyhedrons). The sidewalls of the octahedron are lined with an oxide layer; EDXS and Auger spectroscopy analyses suggested that the oxide is SiO_2 and is approximately 2 nm thick, Figure 3.37b. The void structure with oxide linings in the sidewalls make it unique as a CZ crystal defect, since CZ crystal growth incorporate significant amount of interstitial oxygen. The formation mechanism of the octahedron void defects is not clear. It involves complex interplay between several crystal growing factors. The temperature fluctuations at the growing interface, the growth rate (and axial thermal gradient in the crystal) and dwell time at a critical temperature range (several temperature ranges have been proposed in between 900 and 1100°C [104,105]) in the post-freezing crystal all play a role. Several models have been proposed [106].

3.4 Trends in Large Diameter Silicon Growth

3.4.1 Evolution in Crystal Diameter

As semiconductor technology continues to advance, the IC production is projected to be in sub-60 nm technology generation in 2008 [ITRS]. In parallel with the design rule decrease, the increased circuit design complexity results in an increased chip size. This has been the major driving force for increased wafer diameter for the last 30 years, that is, to increase the required number of IC's per wafer in order to reduce IC manufacturing cost. Figure 3.39 shows the wafer diameter evolution in the industry, since wafer diameter was about 1″ in 1960. The 200 mm was initiated in the late 1980s. In 1995, the

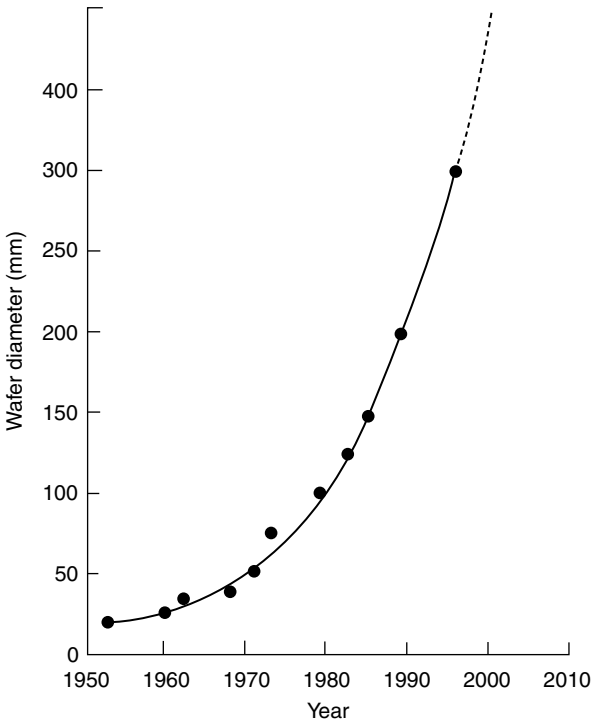


FIGURE 3.39 Evolution of silicon crystal diameter.

development of 300 mm wafer was initiated, targeting for IC manufacture in the 0.25–0.18 μm design rule generation, although the 300 mm era only began in earnest in about 2002 at smaller design rules. Concurrently, Japan launched a project for the development of the 450 mm wafer era technology. In the up-scaling of the wafer diameter, in the 300–450 mm-diameter range, the most significant technical challenges are in the crystal growth. The growing process is far more complex than in the past. Unlike small diameter crystals, the dislocation-free as-grown yield will dominate the cost of the manufacturing of 300–450 mm-diameter wafers. To economically produce large diameter silicon crystals, one needs to employ a large charge size for growing a long crystal. A charge greater than 200 kg for 300 mm-diameter crystal growth or 450 kg for a 450 mm crystal is necessary. At this melt size, the thermal convection is severe. The temperature fluctuations associated with the thermal convection will make initial thin neck growth more difficult. The thermal convection would also result in higher oxygen incorporation in the crystal. It is common to apply an external magnetic field, such as a cusp magnetic field, to the large melt to reduce the thermal convection effect and the melt-crucible interaction. When growing a CZ crystal weighing 150 kg or more employing a thin neck growth to achieve the initial dislocation-free seed, one must consider the risk of fracturing of the thin neck due to the crystal weight exceeding the fracture strength of silicon. An estimate strictly based on fracture strength in tensile mode [107] predicts that the crystal weight limit is about 200 kg when the smallest neck is ∼4 mm in diameter (a targeted neck diameter commonly used for necking). However, for 400–450-mm diameter growth, the crystal weight needs to be in the range of 400–500 kg to be comparable with the 200 mm crystals in production economics. At this weight level and to avoid neck fracture, the "Dash Neck" diameter needs to be larger than 6 mm, a diameter that is difficult to achieve dislocation free structure in the necking process. One of the solutions is to devise a "crystal suspending system" to help support the crystal weight through a "subsidiary cone" grown following the dislocation-free neck is established

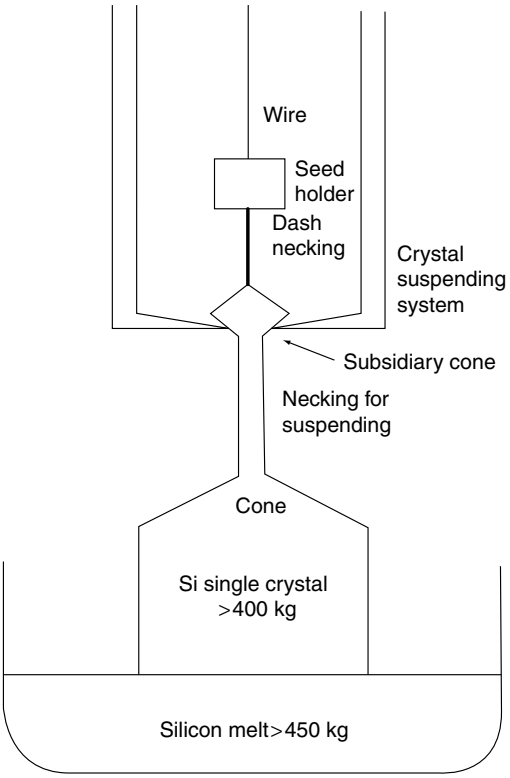


FIGURE 3.40 Schematic of a suspending system for weight support of large diameter growing crystal. (From Yamagishi, H. et al., *Proceedings of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 1996, 59. Reproduced with permission.)

[108]. An example of such a proposed device is shown in Figure 3.40. Another crystal weight related problem is the "creep" phenomenon at the high stress concentration region at the "plastic temperature," $>900°C$. The intersection of the crystal neck and "crown" is such a region [110]. When the stress from the crystal weight (plus the meniscus column and surface tension in the melt) exceeds the critical resolved shear stress for slip, slip dislocations will be generated and propagated down the crystal, along the slip systems, $\langle 110 \rangle / (111)$. Two possible consequences may result. If the slip exits the crystal, the dislocation-free growing process will not be interrupted, but the crystal length above this point will not be useful. If the dislocation reaches the growing solid–liquid interface, the continued growth will not be dislocation-free. The latter case occurs when the crystal length, L, is less than $L = R \tan 54.74°$, where R is the radius of the growing crystal.

Besides weight related problems, the large diameter silicon requires growth rate reduction as well. As the crystal diameter is increased, dissipation of the massive latent heat of solidification from the freezing interface becomes more difficult, since the heat transfer paths are longer. This can be understood from the heat balance shown in Figure 3.3. In silicon crystal growth, a sufficiently high growth rate is essential to maintain a steady crystal growth, in order to maintain the dislocation-free structure. One can enhance the growth rate by enhancing the heat transfer rate via increased crystal surface cooling. Radiation shields have been used to reduce the radiation effect from the melt and the heater [91]. However, by doing so, the thermal gradient is increased resulting in more curved isotherms and interface shapes, enhancing the condition for higher thermal stress. The stress-induced slip can

occur causing structure loss. In the severe case, the high thermal stress can cause crystal cracking. Eventually, the growth rate issue may be a limiting factor in determining the maximum diameter for CZ silicon.

The increase in diameter has a profound effect on the crystal's cooling rate and, therefore, the microdefect formation. If the growing crystal's *V/G* is above the critical value, D defects are generated within part of, or entire radius. It is now understood that the formation of D defects (vacancy clusters), such as COPs, FDPs, and LSTs, the defect density is a function of the dwell time in the temperature range of 900°C–1100°C. [104,105]. Slow cooling reduces GOI defect density. As the crystal diameter is increased, the crystal cooling rate decreases, and the dwell time at 900°C–1100°C increases. It was estimated that for the 300 mm crystal, the dwell time in this temperature range is 50% longer than that of the 200 mm, 100% for the 400 mm case [95]. Therefore, the diameter increase causes the reduction in D defect density. The D defects on wafer surface cause oxide thinning and its density is directly correlated with the defect density in the GOI test. Therefore, it appears that the diameter increase certainly has a positive effect on the microdefect density.

If the diameter increase requires a decrease in the growth rate, then *V/G* may be in the region that the entire crystal is Si interstitial rich. The defects are in the form of small dislocation loops (A and B defects). From the available reports, it appears that the A and B defects have no effect on the GOI, but may cause junction leakage in devices such as charge storage in DRAM, as crystallographic defects, especially if decorated with metallic, are potential recombination centers. The Si self-interstitial-rich silicon also does not favor oxygen precipitation, as SiO_2 induces excess Si atoms in the crystal. Therefore, oxygen precipitation in such silicon may be impeded. This factor plus the fact that the large diameter crystals are low in oxygen (most are grown with MCZ) will require significant research to assure internal gettering in these materials.

It appears that the large diameter crystals grown today and in the foreseeable future will contain either vacancy or self-interstitial type microdefects or both using the "normal" crystal growing processes. However, with the understanding of the relationship between intrinsic point defects and *V/G* during the crystal growth, a growth process may be designed to grow defect-free silicon (Pure silicon or Nearly Perfect Crystal), which is free of COPs and dislocation loops [111]. The schematic in Figure 3.41 shows the relationship between *V/G* and grown-in defect concentration. There is a region of *V/G* within which the concentrations of vacancy and interstitials are below the threshold of defect formation. In order for a crystal to be in the defect-free, a narrow range of *V/G* value has to be maintained across the growing interface during the whole growing process. The vertical thermal gradient can vary from center to edge of the growing crystal, and it is also a function of crystal length grown (or the size of the remaining melt). Therefore, maintaining *V/G* to a target value presents major challenges in process design as well as the design of grower's hot-zone thermal characteristics. With the availability of defect-free silicon, via growing process control, the device makers have more options in materials selection based on factors related to cost, nature of the processing, etc. The epitaxial wafers and hydrogen (or argon) annealed wafers [113] are also widely used. The current hydrogen/argon annealing process can provide a 5–10 μm deep defect-free zone and, with designed thermal ramping, can provide nuclei for internal gettering (IG; by temperature ramping during anneal). The epitaxial silicon layer has a higher quality than the melt grown bulk silicon. The layer is essential free of interstitial oxygen, carbon, and microdefects, and lower in surface particle and metal contamination than the bulk wafer. The DRAM manufacturing group has been the largest user of the polished wafers. The microdefect problem associated with the bulk wafer may drive a significant fraction of the DRAM manufacturers to switch to epitaxial wafers. Two possible epi structures exist: p/p^+ and p^-/p^-. The former has also been popular with the microprocessor and application specific integrated circuit (ASIC) manufacturers, the advantages including improved gate oxide quality, internal gettering, latch-up immunity, etc. The p^-/p^- approach is useful to "mask" the microdefect problems in the polished p^- wafers. The microdefects, neither dislocation loops nor voids are not found to extend into the epitaxial layer during the epitaxial growth [114].

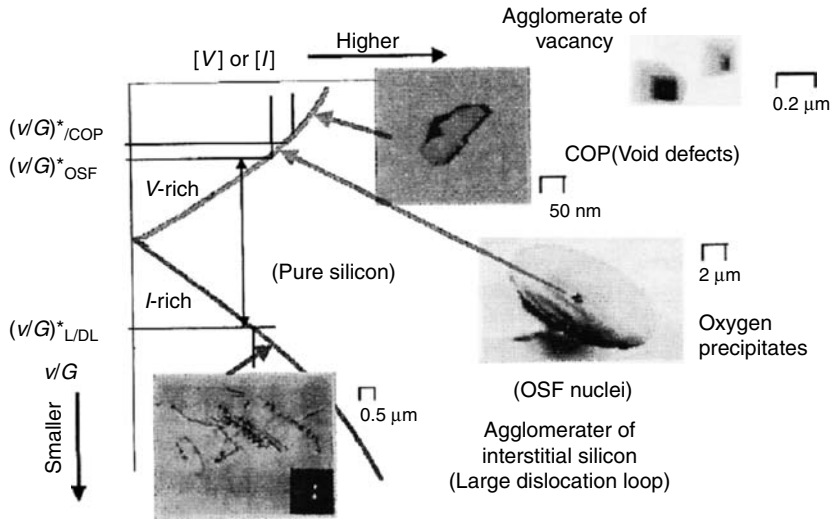


FIGURE 3.41 Schematic showing grown-in defects depending on the *V/G* ratio, where *V* is growth rate and *G* is the thermal gradient at the solid/melt interface. (From Rozgonyi, G. A., *Semiconductor Silicon 2002*, ed. Huff, H. R., Fabry, L., and Kishino, S., Electrochemical Society, Pennington, NJ, 2002, 149. Reproduced with permission from Electrochemical Society.)

3.4.2 Continuous Czochralski Silicon Growth

The idea of continuous CZ growth is due to the fact that, as the silicon diameter continues to increase the maximum grown crystal length of the batch CZ process is limited by the charge size. The initial motivation of the approach of the continuous growth was for the increased crystal length and thus improved throughput and operation cost of the CZ grower. A continuous CZ growth based on a two-container system was first demonstrated by Fiegl [24]. In addition to increased crystal length, many other desirable crystal properties were also demonstrated, such as improved axial uniformity of both

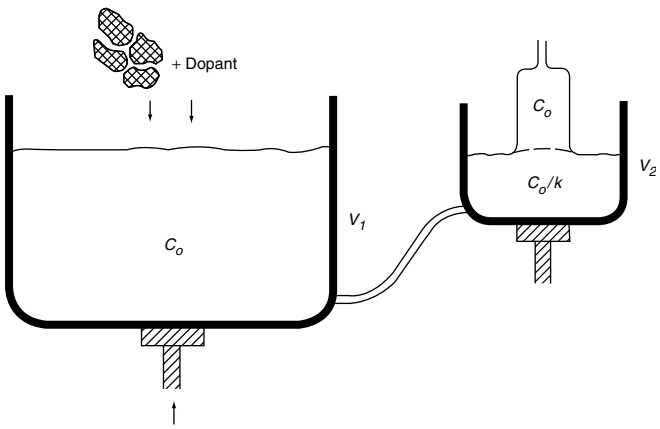


FIGURE 3.42 Schematic showing a generalized two-container arrangement for crystal growth from a constant volume melt by continuous feed. (From Lin, W. and Benson, K. E., *Annual Review of Materials Science*, 17, 273, 1987. Reproduced with permission from Annual Reviews.)

dopant and oxygen concentrations. In particular, the arrangement affords the use of a small and shallow-melt CZ set-up in the growing container. In many ways, the two-container arrangement with crystal pulling from a constant melt volume is the same as the double-crucible operation. Figure 3.42 is a generalized two-container system reconfigured from a constant-volume double-crucible, where V_1 and V_2 are the melt volumes in the outer and inner crucibles, respectively, in a double-crucible set-up. The reconfigured arrangement facilitates the addition of polysilicon and dopant. Only the outer crucible (the feeder) needs to be lifted or adjusted for the melt level control for constant volume/melt level in the growing crucible. The system offers the same advantages as the double crucible in an axial uniformity in dopant and oxygen and the benefit of a small melt effect. Figure 3.43 shows the uniform axial oxygen profile obtained due to "constant volume" growth. The uniform axial dopant profile at concentration C_0 is obtained when the initial doping of the growing crucible is C_0/k and the feeding crucible concentration is maintained at C_0. The incorporation behavior of impurities in the feed material is similar to one-pass

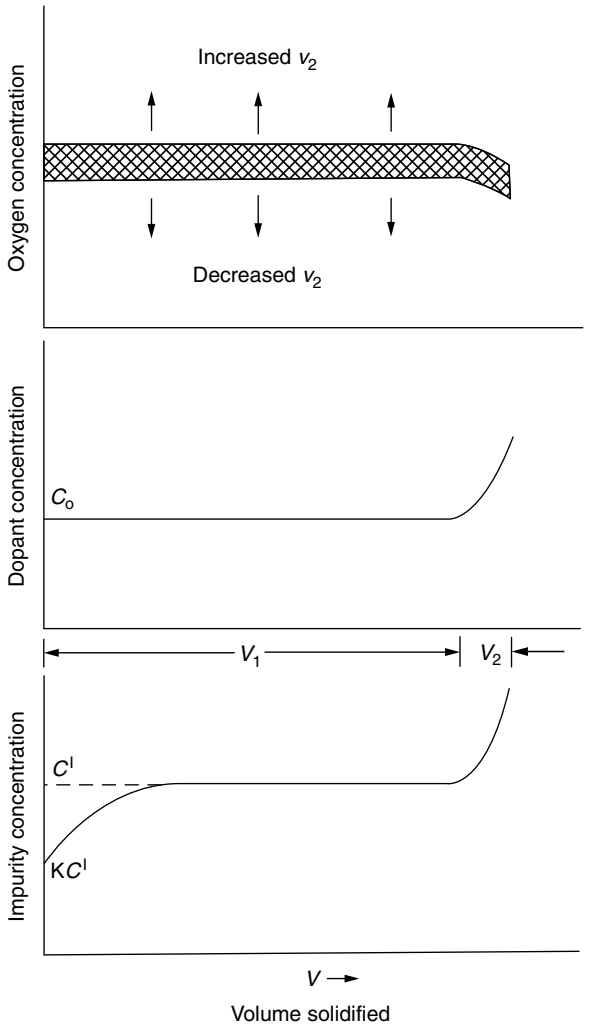


FIGURE 3.43 Axial oxygen, dopant and impurity concentrations along the crystal grown from the arrangement shown in Figure 3.43. The oxygen concentration level can be adjusted by changing the melt volume (From Lin, W. and Benson, K. E., *Annual Review of Materials Science*, 17, 273, 1987. Reproduced with permission from Annual Reviews.)

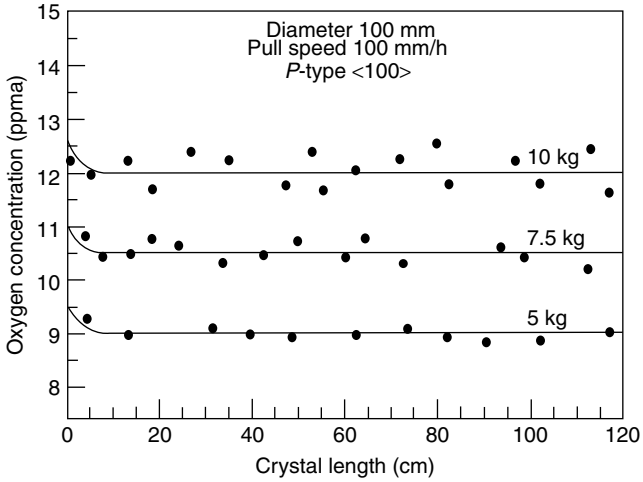


FIGURE 3.44 Uniform axial oxygen distributions from "constant volume" continuous-feed silicon growth. The oxygen level is shown to depend on the melt size. (From Fiegl, G., *Solid State Technol.*, August permission from Solid State Technology.)

zone leveling [22] with the "zone width" being equivalent to the melt volume in the growing crucible, V_2. In the continuous growth mode, V_1 is the total melt volume passing from the feeding (outer) crucible to the growing crucible before V_2 is consumed and reduced. The oxygen concentration level is determined by the volume and aspect ratio of the melt in the growing crucible (see Figure 3.44).

While the continuous "feed and pull" system with two containers appears straightforward, many engineering challenges remain to be solved or improved for the method to be practical. Among the problems to be solved are the liquid melt transfer and establishment of thermal stability and radial thermal symmetry in the melt while receiving replenishment melt from an external source. In recent years, with the availability of high-purity, small-diameter silicon beads (~0.1–1 mm) from the fluidized bed process, "feed and pull" may be carried out in a double-crucible arrangement as illustrated in Figure 3.45a and b. In case of configuration in Figure 3.45a, the partition between the inner and outer crucibles can be inserted, after the polysilicon charge is completely melted, before the seeding begins [115]. Crystal length of 2 m grown under such continuous "feed and pull" mode was demonstrated. If the total melt volume is kept constant, the partition separating both the melt concentrations becomes

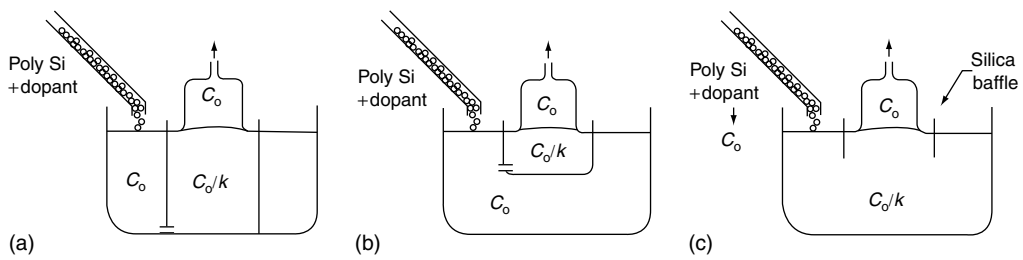


FIGURE 3.45 Crystal growth from double-crucible arrangements with (a) constant melt level or (b) constant inner melt volume, maintained by continuous-feed (c) crystal growth from a single container equipped with a circular silica baffle. Melt level is keep constant by continuous feed. The melt concentration is maintained at C_0/k. (From Lin, W. and Benson, K. E., *Annual Review of Materials Science*, 17, 273, 1987. Reproduced with permission from Annual Reviews.)

unnecessary. The resulting "one melt" growth retains the advantages of double-crucible arrangements, Figure 3.45c. In this arrangement, the oxygen concentration level is controlled, as in the two-container case, by the melt volume and aspect-ratio. However, the application of forced convection for an additional oxygen incorporation control is more restricted than for standard batch processes. Furthermore, in this design, the single crucible contains a physical barrier to prevent unmelted silicon particles from reaching the growing crystal and a baffle that reduces thermal convection. One may also view the continuous feed mechanism as playing the role of the outer crucible, which supplies silicon with dopant concentration C_0 (the intended concentration in the crystal). The small-diameter polysilicon beads add a new dimension to the development of continuous "feed-and-pull" silicon growth. Shiraish et al. [116] uses liquid-feed for continuous crystal growth of large diameter silicon (150 and 200 mm-diameter crystals). In this approach, the polysilicon rods are melted into liquid silicon immediately above the growing melt, inside the growing chamber, and continuously fed into the CZ melt. The continuous-mode growth provides flexibility where the melt volume and aspect ratio of the growing crucible can be adjusted for oxygen incorporation level. This is especially important for the low oxygen incorporation, which cannot be easily attained in the standard CZ growth with a melt size of 80 kg or larger.

3.5 Wafer Preparation

Silicon semiconductor devices are mostly fabricated on polished wafer or epitaxial wafer. Thus, the first step in device fabrication is the preparation of mirror polished, clean, and damage-free silicon surfaces in accordance with the specifications. As the design rule of device fabrication advances into the deep sub-micron region, the device processing and performance are more sensitive to the starting material's characteristics. The requirements of the geometrical tolerance of the polished wafers as well as their bulk characteristics are becoming more stringent. The polished wafers are prepared through the complex sequence of shaping, polishing, and cleaning steps after a single crystal ingot is grown. Although the detailed shaping processes vary depending on the manufacturer. The processes described below are generic in nature. Newly introduced processing technologies will be discussed where appropriate. Figure 3.46 is a flow chart showing a generic wafer shaping process.

The single crystal ingot is first evaluated for crystal perfection and resistivity, before it is surface ground to a cylindrical shape of a precise diameter. Flat(s) or a notch with preferred crystallographic orientations are ground on the ingot surface parallel to the crystal axis. The primary flat or notch, for example, is positioned perpendicular to a ⟨110⟩ direction on a (100) wafer and is used for alignment of the wafer in the device processing with automated handling equipment. The primary flat, or notch, also serves as an orientation reference for chip layout, since devices fabricated on wafers are crystallographically oriented. The existence of secondary flat on the wafer, shorter than the primary, is used to identify the wafer surface orientation and conductivity type [117].

3.5.1 Slicing

The slicing operation produces silicon slices from the ground ingot. Slicing defines the critical mechanical aspects of a wafer, such as thickness, taper, warp, etc. The slicing is commonly carried out by an inner diameter (ID) circular saw (Figure 3.47a), after the ingot is rigidly mounted to maintain an accurate crystallographic orientation as previously determined by x-ray diffraction. The ID saw uses a thin stainless steel blade bonded with diamond particles on the inner edge of the blade. Recently, the development of multiple-wire saws has enabled the silicon slicing to result in high throughput and superior mechanical properties such as significantly reduced bow and wrap. Figure 3.48b shows a schematic of a multiple-wire saw. In this arrangement, parallel, equally spaced, and properly tensioned stainless steel wires spun across two pulleys are part of a single stainless steel wire winding through a complex set of pulleys. Cutting of multiple slices results when the ingot is pressed against the traveling wires under injection of slurry. Although the cutting rate is much slower than the ordinary ID saw (ordinary ID saw is 80–100 times higher

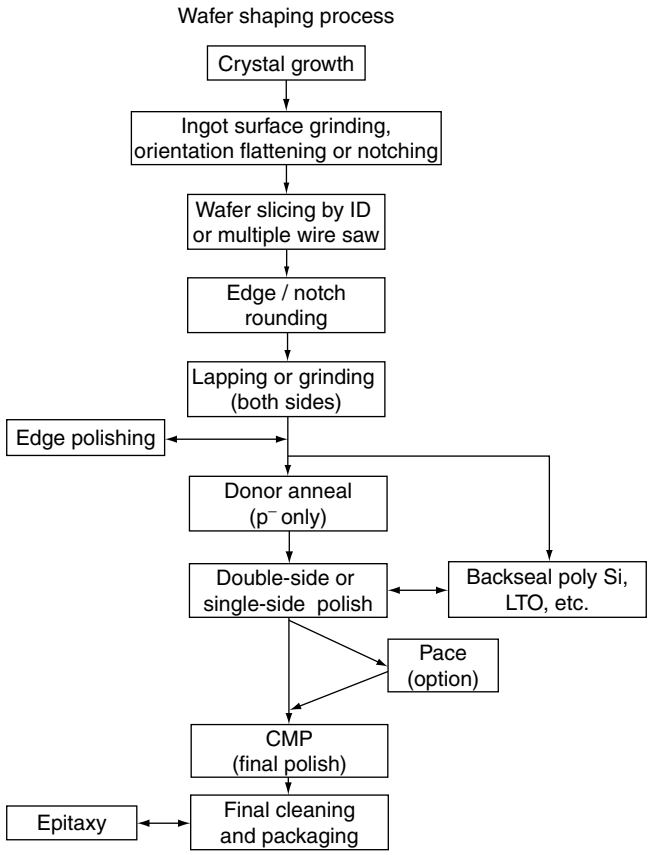


FIGURE 3.46 Flow chart describing generic steps involved in wafer preparation employing modern technologies.

rate), as many as 300 slices can be produced simultaneously. Besides higher throughput, the multiple-wire saw has other major advantages over the ID saw. Slicing by the multi-wire is actually the result of low speed grinding/lapping action by the slurry. Improved bow, warp, TTV, and taper are much easily obtained than with the ID saw. In addition, the slow lapping action by the moving wire results in small kerf loss, which affords more slices per inch of ingot. It is shown that the kerf size loss is very close to the diameter of the wire used. The multiple-wire saw offers material savings (reduces kerf loss by 30% compared to the ID saw), increased productivity, and improved wafer mechanical properties. It has been used for slicing 200- and 300-mm-diameter wafers and is expected to be used for the future "diameter generations."

3.5.2 Chemical Etching

Chemical etching of the slices is done to remove mechanical damage induced during the previous shaping steps-ingot surface grinding and slicing. The etching can be carried out by either an acidic solution or a caustic etchant. The acidic system is mostly based on HNO_3–HF system (or with modifiers such as acetic acid [118]). The surface material removal is the result of two-step reactions. The Si surface is first oxidized by HNO_3 to form SiO_2, followed by its removal by HF. The acid etch produces a smooth and shiny surface. However, since the reaction is exothermic, temperature control is critical in order to maintain uniform etching. Caustic etching uses a alkaline solution [119], such as KOH, with certain stabilizers. The KOH etch offers a uniform etching rate, but produces a rougher surface than the acid

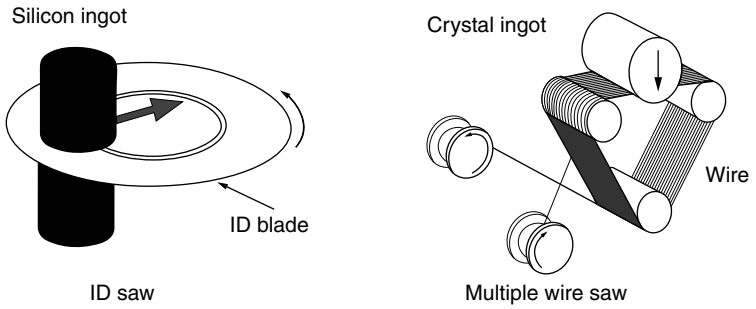


FIGURE 3.47 Schematics showing the traditional inner diameter (ID) saw and recently developed multiple-wire saw for silicon wafer slicing.

etch, since KOH etching rate is crystallographic orientation dependent. Chemical etching of the slice may be repeated after subsequent mechanical operations, such as edge rounding and lapping/surface grinding to remove mechanical damage.

3.5.3 Edge Rounding

The square edge of sliced silicon wafers is rounded by an edge grinder. The rounded-edge wafers greatly reduce mechanical defects, such as edge chips and cracks induced by wafer handling. Edge chips and cracks can serve as stress raisers, which facilitate the onset of wafer breakage or plastic deformation, and slip dislocations during thermal processing. In addition, the rounded edge eliminates the occurrence of epitaxial crown (thicker epitaxial layer at the wafer edge) in the epitaxial deposition process and pile-up of the photoresist at the wafer edge. The shape of the rounded edge usually follows an industrial standard (i.e., SEMI standard) in which the edge profile fits within the boundary of a standard template. However, variations from the "standard" exist. In some applications, the rounded edge is modified to be more "blunt" in order to facilitate the chemical–mechanical polishing (CMP) operation for inter-level dielectric planerization. In this case, the "blunt" edge supposedly prevents the wafer from slipping out of the template during polishing. Other applications require more "rounded" edge shapes for increased strength. Often, a compromise on the edge profile is required.

3.5.4 Lapping/Grinding

The lapping of the silicon slice surface takes place when the slice is ground between two counter rotating cast iron plates in the presence of an abrasive slurry, usually a mixture of micron-sized alumina or silicon carbide particles suspended in a solution. The purpose of lapping operation is to remove the non-uniform damage left by slicing, and to attain a high degree of parallelism and flatness, both global and local. In fact, post-lapping slices possess the best mechanical characteristics in the entire shaping process flow. The subsequent mirror-polishing operation generally degrades the flatness characteristics attained by the lapping operation. However, lapping with slurry also introduces fresh damage to the silicon surface, which requires subsequent chemical etching and chemical–mechanical polishing for removal. Chemical etching and CMP of the silicon surface degrades the wafer flatness. To circumvent this situation, surface grounder (with a precision grinder bonded with diamond particles) on both sides of the wafer is employed to achieve surface flatness with reduced surface damage. With the reduced surface damage, the need for chemical etching and CMP is also reduced and good mechanical properties may be retained.

3.5.5 Polishing

Polishing is accomplished by a chemical–mechanical polishing process involving a polishing pad and a slurry. The polishing slurry is usually an alkaline colloidal solution containing micron-sized silica particles. While CMP is used to remove surface damage and to produce a mirror-finished surface, it also degrades the wafer flatness achieved by lapping/grounding. Therefore, it is essential to optimize operational parameters so as to minimize the polishing time and flatness degradation. Double-side polishing (CMP on both front and back surfaces simultaneously) has been found to result in superior flatness than the single side polishing arrangements. The combination of surface grinding (on both the sides of the wafer) and double-side CMP has shown to result in a superior total indicator reading (TIR), total thickness variation (TTV) and local flatness. Such an approach is becoming a standard manufacturing process for large diameter wafers ($\geq$300 mm) preparation.

The wafer preparation via mechanical methods (i.e., grinding, CMP etc.,) has its limits in the degree of flatness that it can achieve. To supplement and to fine tune the local topography for further improvement in local flatness, tools such as plasma assisted chemical etching (PACE) [120] have been developed. Such a tool employs a spatially confined plasma with a scanning mechanism to allow material removal to be controlled as desired over the wafer surface. The PACE utilizes low energy neutral ions (i.e., $\ll$1 eV) rather than energetic ions, which are involved in reactive ion etching. Therefore, PACE produces minimum or no subsurface damage.

3.5.6 Cleaning

Wafer surface contamination can affect electronic device performance. The contaminants can be attached to the wafer surface physically or chemically. Cleaning of the wafers is necessary at many steps in device fabrication processes, as well as during the wafer shaping processes. The silicon wafer must be free of contamination before it is shipped to the device fabrication line. The cleaning process for the removal of surface contaminants during wafer shaping and polishing processes is discussed below.

In general, the contaminants can be classified as molecular, ionic, or atomic. Typical molecular contaminants include waxes, resins, and oil used in polishing and sawing operations, and material from the plastic containers used for slice transport and storage. Molecular contaminants are absorbed on the wafer surface by weak electrostatic forces. They should be removed before subsequent cleaning involving chemical reactions. The ionic contaminants, such as Na^+ Cl^-, and F^- are present after wafer treatments in HF-containing or caustic solutions. They are attached to the wafer surface by chemical absorption. The atomic contaminants of concern are due to the transition metal atoms, such as Fe, Ni, and Cu. The transition metals and ionic species can cause degradation in device performance.

Chemical cleaning is an effective method to remove contaminants on the wafer surface. Many chemical cleaning processes have been developed. The process that is widely used in the semiconductor industry is the so-called "RCA Clean" [121], which consists of two consecutive cleaning solutions, including H_2O–H_2O_2–HN_4OH (Standard clean 1, SC1) and H_2O–H_2O_2–HCl (Standard clean 2, SC2). The SC1 clean, with volume ratios typically 5:1:1 is to remove organic contaminants by both the solvating action of NH_4OH and strong oxidizing action of H_2O_2. The NH_4OH can also form soluble complexes with some metals such as gold, copper, nickel, and cobalt. The SC2 clean, with typical volume ratios of 6:1:1 removes transition and alkali metals from wafer surface, and prevents redeposition from the solution by forming soluble metal complexes (with Cl^-). The SC1 can also remove particles physically attached to wafer surface by etching effect of NH_4OH which detaches the particles from the wafer surface. The repulsion effect of the opposite charges transferred from the electrolyte NH_4^+ and OH^- to the wafer surface and detached particle, respectively, prevent the particles from redepositing on the wafer surface. A modified RCA Clean [122], by adding a brief etch in diluted HF solution after SC1 was designed to eliminate the thin oxide layer grown on silicon surface due to the SC1 process. The thin oxide resulted from the SC1 was thought to hinder surface for cleaning by the SC2. Many modifications of the "RCA clean" exist (published and unpublished). Most of the modifications are on the volume ratios. For

example, in order to reduce silicon surface roughness, there is a trend to greatly reduce the volume fraction of the NH_4OH in SC1 from the original formula. The effect of surface roughness of silicon on the gate oxide integrity has been reported as significant when the oxide thickness is thinner than 5 nm, although the literature is often not consistent, probably due to different process conditions. Other chemical cleaning solutions/processes, such as piranah (H_2SO_4–H_2O_2–H_2O), ozonated water, etc., have also been shown to be effective. However, the "RCA clean" and its modified versions have been the most popular cleaning processes used by the semiconductor silicon manufacturers.

3.6 Epitaxial Growth

3.6.1 Silicon Epitaxial Wafer

An epitaxial silicon wafer refers to a structure where an epitaxial layer is grown on a single crystal silicon substrate by chemical vapor deposition (CVD), normally at a high temperature. The CVD process usually involves the hydrogen reduction of high purity silicon tetrachloride ($SiCl_4$), trichlorosilane ($SiHCl_3$), or dichlorosilane ($SiHCl_2$) to form solid silicon. An added source gas in the reaction, such as diborane (B_2H_2) or arsine (AsH_3) provides dopant atoms for *n*- or *p*-type electrical carriers, respectively, in the epitaxial layer. The primary purpose of the epitaxial growth is to create a layer with a different, usually lighter, concentration of electrically active dopant than the substrate. Depending on the IC characteristics, the epi layer must meet a set of specifications for thickness, electrically active dopant concentration, sharpness of the epi-substrate interface, defect density, and contamination.

The epitaxial layer was initially applied to bipolar devices. Typically, a lightly doped layer is grown over a substrate containing a pattern (often referred to as sub-collector) of opposite type impurity by diffusion or ion implantation. The structure allows a vertical transistor to be built with a minimum of collector resistance and readily permits junction isolation between devices (see Figure 3.48). The most common application of the epitaxial layer in CMOS device processing involves a lightly doped layer over a heavily doped substrate of the same conductivity type, i.e., either p/p^+ or n/n^+ structure. The p/p^+ is by far the dominant universal epitaxial structure. It has been used by CMOS-based logic, microprocessor, ASIC, and some DRAM manufacturers. Figure 3.49 is a schematic of a CMOS structure built on p^- layer of the p/p^+ wafer.

It is realized today that p/p^+ wafer has many advantages over the bulk p^- wafers. The initial motivation for using such a structure was to reduce the metal–oxide–silicon (MOS) device's leakage current. Since a lightly doped bulk substrate has a higher concentration of minority carriers, the carriers can diffuse hundreds of micrometers to space-charge layers and are collected as reverse-bias leakage currents. This minority-carrier diffusion current can dominate over leakage current generated within the space-charge layers, especially at higher operating temperatures ($\geq 40°C$) [123]. A technique that can circumvent this problem is to form a p^- layer ($\sim 10^{15}$ atoms/cm^3) as an epi layer grown on a heavily doped substrate ($\sim 10^{19}$ atoms/cm^3) [124,125]. The p^+ substrate has few minority carriers (electrons), so the minority carriers are only generated from the thin epi layer. Thus, the diffusion current

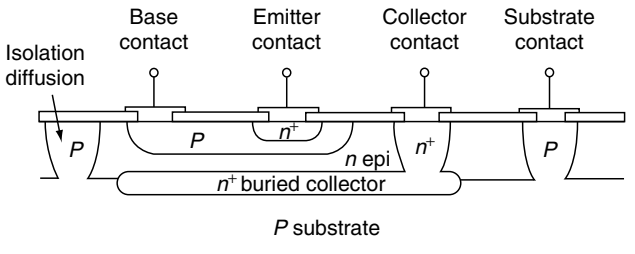


FIGURE 3.48 The role of epi layer in a bipolar device.

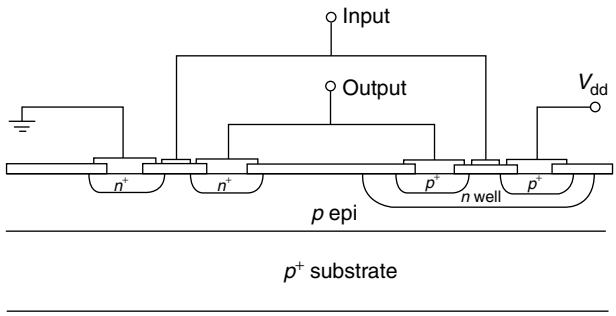


FIGURE 3.49 The role of epi layer in a complementary MOS device.

from the substrate is suppressed, even though the minority-carrier diffusion lengths (in the epitaxial layer) are long; the minority-carrier diffusion length is small in the p+ substrate. This is especially important in preserving holding times in dynamic nodes (e.g., DRAM) [123].

Since its initial application in CMOS, many added beneficial functions have been found to be associated with the p/p^+ structure. The major ones include its effects in minimizing soft errors, preventing latch-up of the device, and providing sites gettering of harmful impurities. From the device design view point, using p/p^+ structure for CMOS fabrications is the best solution for avoiding latch-up without resorting to circuit design modifications, such as adding "guard rings," etc. The latch-up is due to the turn-on of a parasitic 4-terminal, n–p–n–p lateral transistor between neighboring NMOS and PMOS transistors, as shown in Figure 3.49. However, if the formation of a vertical parasitic p–n–p is favorable with the presence of p^+ substrate, the turn-on of the lateral 4-terminal device may be suppressed. In a first approximation, in order to suppress the lateral p–n–p–n device, the effective epi thickness (taking into account the boron up-diffusion after device processing) must be smaller than the separation of two neighboring n–p transistors. Latch-up simulators are available to provide information for designing latch-up-free design layout, from which an appropriate epi thickness may be estimated.

p^+ Substrates in p/p^+ wafers have been found to serve as an excellent intrinsic gettering sites. Two effects have been observed. The first is the gettering effect due to p^+ silicon's high efficiency in generating oxygen precipitates–dislocation complexes. The oxygen precipitation in p^+, with resistivity around 10 m ohm–cm, has fast kinetics and can result in a precipitate density, which is more than an order of magnitude higher than in p^- (10 ohm–cm) under a Lo–Hi annealing condition [41]. Much of the oxygen precipitation behavior is discussed in Section 3.7. The second gettering effect by the p^+ silicon is due to the "segregation" effect, which drives metallic impurities to the p^+ region from the p^- epi layer under a "segregation" annealing [126]. The p^+, like other highly doped (degenerately doped) silicon regions, such as diffused phosphorus layers, source/drain regions in the CMOS structure, have long been observed to getter metal impurities and perceived as due to an enhanced solubility effect. The segregation effect of the impurities (such as Fe) was proposed to be due to the difference in the Fermi levels of p^+ and p^- silicon [127].

One of the unique features associated with the p/p^+ structure is the possible existence of misfit dislocations at the epi-substrate interface (see Figure 3.50). These dislocations can act as gettering sites as well. It is well known that doping with boron causes the silicon crystal lattice to contract ($\sim$0.014 A/atom% boron added) [128]. The vast difference in the boron doping levels in p/p^+ ($\sim 10^{19}$ atoms/cm^3 in the substrate while the epi layer is doped with 10^{15} atoms/cm^3) results in a lattice mismatch between the epitaxial layer and the substrate during CVD deposition [129]. Under this condition, the epitaxial growth is accompanied by lattice stress, tensile on the epi layer side, and compressive on the substrate side. The misfit stress increases as the layer continues to grow, until the local stress exceeds the elastic limit at the deposition temperature, at which time the stress is partially relaxed by forming misfit dislocations at the epi-substrate interface. The onset of misfit dislocation formation during epi deposition is a

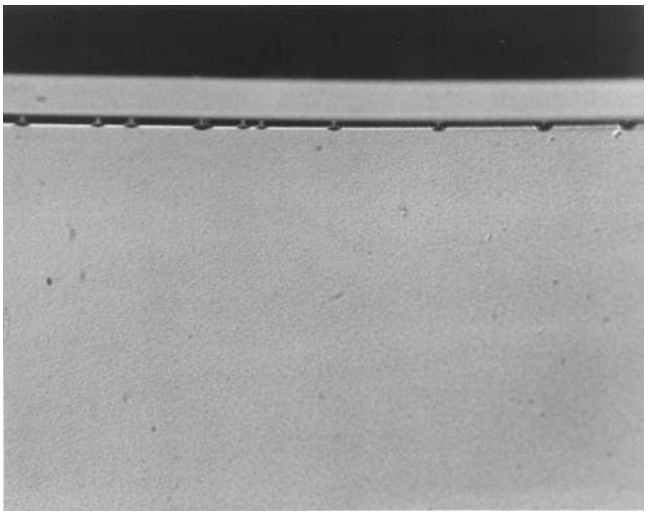


FIGURE 3.50 Cross-sectional view of an etched p/p^+ epi wafer, showing etched misfit dislocation pits at the epi-substrate interface.

function of the degree of lattice mismatch, deposition temperature, and deposition thickness. The unrelaxed misfit stress will contribute to both the wafer bow and the wrap. It should be reminded that the misfit dislocations formed on the (100) p/p^+ epi structure are contained in the (100) interface, with a Burger's vector of the form $b = a/2\langle 110\rangle$ [130]. Since the (100) plane is not a glide plane in the silicon lattice, the dislocation is immobile; it is "locked" [131] in the interface. Furthermore, the dissociation of such a dislocation into two mobile dislocations in two inclined {111} slip planes is energetically unfavorable. However, it is possible to move misfit dislocations in silicon by a "climb" mechanism under high stress conditions. Experimental evidences [129] show that the movement is towards the substrate side, i.e., the side with the smaller lattice constant. The misfit dislocations so formed are stable and could contribute to the gettering effect. They present no harmful effect to the epi layer structure.

Perhaps the most fundamental difference between the polished bulk wafer and p^- layer of the p/p^+ wafer is that CVD silicon is superior in quality. The epi layer is free of oxygen and carbon incorporation. The incorporation of dopant is uniform and free of local fluctuations (such as striations in CZ materials discussed in Section 3.3.2.4.2). In general, CVD epitaxial silicon has been recognized as superior materials for GOI in device processing than bulk CZ silicon (see Figure 3.51). More significantly, in deep sub micron design-rule technologies, CVD silicon epitaxial material is of vital importance and beneficial to gate oxide integrity. Chemical vapor deposition epitaxial silicon layer is also free of grown-in microdefects stemming from the agglomerations of point defects encountered in the melt grown silicon (CZ). Grown-in "D" defect (such as COPs, LSTs, and FDPs) density in CZ silicon has been correlated with the oxide defect density. The issue of grown-in defects is discussed in Section 3.3.3. However, the defects due to agglomeration of point defects in CZ growth tend not to extend into the epitaxial layer during CVD deposition. It has been demonstrated that as thin as 0.3 μm epitaxial layer, grown on a CZ substrate containing COPs can mask the harmful effect of COPs [114]. In consideration of the impact of grown-in microdefects on device yield and reliability, many device manufacturers will be prompted to switch from bulk CZ wafers to p/p^+ or even p^-/p^- epitaxial wafers.

3.6.2 Heteroepitaxy

Heteroepitaxial growth refers to epi deposition, where the deposited layer and substrate's lattice constants are slightly different due to difference in chemical composition, although they may be of the same crystal

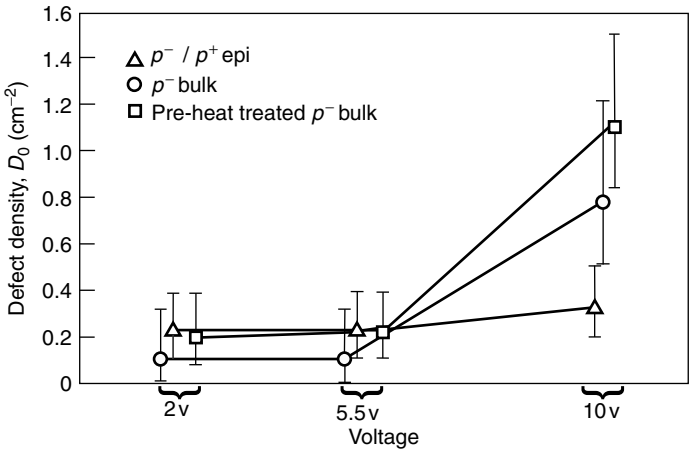


FIGURE 3.51 Defect density (D_0) extracted from voltage measurements on thin oxide grown on three different silicon surfaces. (From Boyko, K. C., Feiller R. L., and Lin, W. Unpublished.)

system. Heteroepitaxy has been employed in the fabrication of new emerging Si materials and advanced device processing. For example, strained silicon growth, SiGe alloy deposition for recessed S/D, SiGe for a heterojunction bipolar transistor (HBT) via selective epitaxy, are among the recent applications. In heteroepitaxy, due to a difference in crystal lattice constant across the interface, there exist a strain in the grown epitaxial layer in order to accommodate the lattice mismatch. The layer is pseudomorphic when the layer thickness is thinner than a certain critical thickness [132], above which the stress in the film is partially relaxed by the formation of dislocations. Figure 3.52 shows a lattice model for strained pseudomorphic epi layers in two different heteroepitaxy arrangements. In both cases, the epi thickness is thinner than its critical thickness. In Figure 3.52a situation, the epi layer is in a state of biaxial compressive strain, and in Figure 3.52b, a state of biaxial tensile strain.

When the pseudomorphic layer continues to grow in thickness and reaches a certain critical thickness [132], the epi layer partially plastically relaxes and releases the lattice strain by generation of misfit dislocations at the growing interface. This is shown in Figure 3.53. The completely relaxed heteoroepitaxial layer is in equilibrium state. As will be discussed in the following sections, both pseudomorphic and relaxed epi layers have their applications.

3.6.3 Selective Epitaxial Growth

Compared to blanket silicon epitaxy discussed above, selective epitaxy growth (SEG) has not been widely used. Silicon bipolar transistors have used, selective epi other than SiGe for improved performance. A 1988 IBM paper [133] reports on a selective epitaxy base transistor, which forms the base with boron-doped selective epi rather than boron implantation. The development of the SiGe HBT made use of the SEG in bipolar and BICMOS fabrication in the 1990s. The SEG has also been used to make the source and drain (S/D) elevation for CMOS transistors to solve the problem on silicon real estate limitation. The process becomes indispensable with SOI processing. More recently, SEG has been employed for growing recessed S/D with SiGe to induce uniaxial compression strain in the PMOS channel.

The SEG of silicon has been studied from the early days of silicon technology [134–136]. In SEG of silicon, epitaxial deposition only takes place on "windows" of bare silicon, single crystal substrate of a patterned wafer. Nucleation is suppressed elsewhere in regions that is covered by silicon dioxide (or silicon nitride). The major difference in SEG process from that of conventional epitaxial growth is the addition of extra HCl to the SiH_2–H_2 or SiH_2Cl_2–H_2 chemistries, although the by-product of the

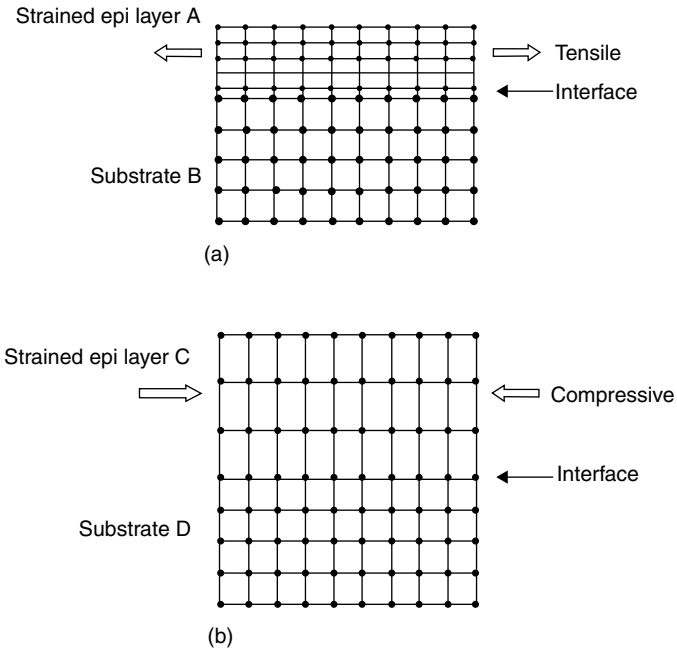


FIGURE 3.52 Lattice model of strained pseudomorphic layers grown with (a) Lattice A < Lattice B, and (b) Lattice C > Lattice D. Layer A is in biaxial tensile strain and layer C is in biaxial compressive strain. Layer A and Layer C are thinner than their respective "critical thickness."

chlorine-containing source gas also provides the selectivity. Figure 3.54 illustrates the SEG on a patterned wafer. A SEG process at a reduced pressure and a high temperature along with a careful pre-deposition clean (such as high temperature H_2 bake) can result in a grown silicon layer with minimum crystalline defects. When doing SEG on (100) substrates, {311} facet growths are found to be common at the edges of the pattern [136]. When the edges of the pattern are aligned in the ⟨100⟩ directions, {311} facets will be situated in the corners, and their size and effect are minimized.

3.6.3.1 SEG Growth of SiGe for HBT

The epitaxial growth of SiGe is the heart of the heterojunction NPN transistor technology. The modern SiGe HBT was developed in the 1990s. A SiGe HBT is similar to a conventional Si bipolar NPN transistor

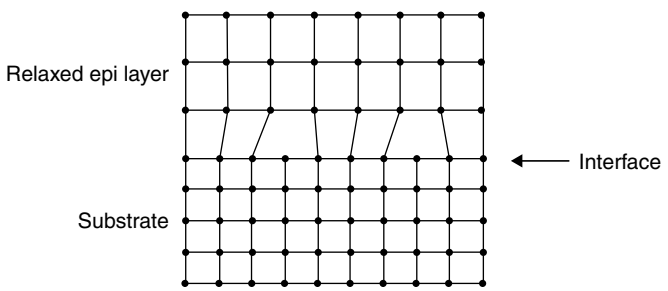


FIGURE 3.53 Lattice model of heteroepitaxy after strain relaxation accompanied by generation of misfit dislocations. The relaxed epi layer is thicker than the critical thickness.

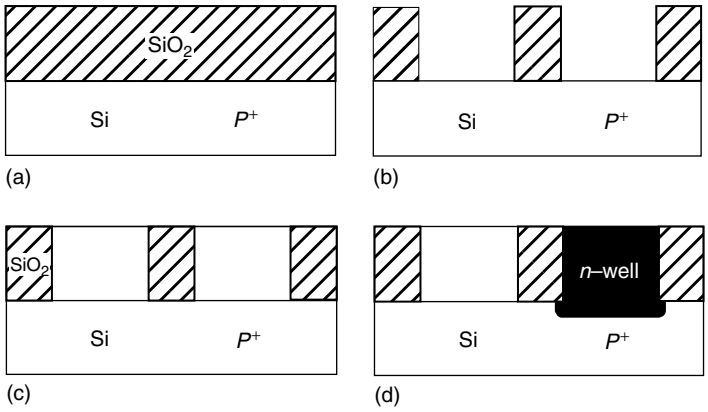


FIGURE 3.54 Schematic illustration of a typical selective epitaxy growth (SEG) process for device isolation: (a) oxide deposition; (b) window formation; (c) epi growth (d) n-well drive-in. (From Borland, J. O. and Drowley, C. I., *Solid State Technol.*, 28, August, 1985, 141. Reproduced with permission from Solid State Technology.)

except for the base. SiGe, a material with narrower bandgap than Si, is used as the base material. The Ge composition is typically graded across the base, with the Ge decreasing from the collector side to the emitter side (peak concentration ranging from 10 to 25 atom% Ge, depending on the application). This creates an accelerating electric field (sloped conduction band) for minority carriers moving across the base, as schematically shown in Figure 3.55. A direct result of the Ge grading in the base is higher speed, and thus higher operating frequency, $fT \sim 100$ GHz. Typically, thin SiGe a few tens of nanometer thick is deposited selectively on the substrate (collector) at a moderate temperature (650°C–750°C) and a reduced pressure of a few torr using SiH_2Cl_2, GeH_4, B_2H_2, and HCl. The boron doping level of the layer is

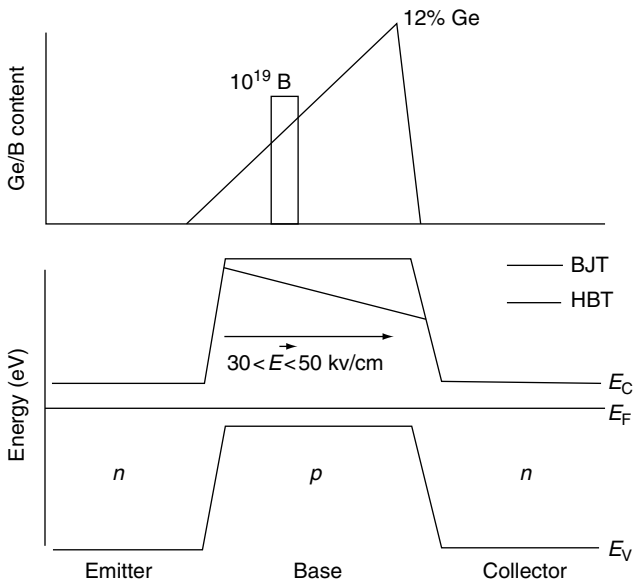


FIGURE 3.55 Schematic showing effect of Ge Grading across the base of a 300-Angstrom-basewidth NPN bipolar transistor, which results in a large built-in pseudopotential, greatly reducing base transit times for electrons. (From Meyerson, B. S., *IBM J. Res. Dev.*, 44, 2000. Reproduced with permission from IBM.)

in the order of 1E18/cm^3. The SEG can be carried out in a single wafer RTCVD reactor. The higher speed HBT would require higer Ge doping, 30% or higher, and a higher boron doping greater than 1E19/cm^3 in order to keep the intrinsic base resistance low.

Since the atomic radii of Ge is ~4% larger than that of Si, the doping of Ge (to substitute Si) causes the lattice parameter to increase linearly with the Ge concentration. There is significant lattice mismatch and strain when depositing SiGe on Si substrate. For SiGe HBT applications, it is important to realize that the SiGe layer, typically several hundred angstroms in thickness, so deposited is below the critical thickness for the Ge concentration of interest. Therefore, the layer is under elastic strain and is not relaxed. It is in a pseudomorphic state as discussed above. Since there is no plastic relaxation, there is no misfit dislocation formation at the interface. The SiGe is in biaxial compressive strain since its lattice parameter is larger than the substrate, Si. This is shown in Figure 3.52b.

While the Ge atom is larger than Si, the doping of Ge causes the Si lattice to expand by 0.0022 Angstrom/atom% of Ge [138], but doping by the smaller boron atom would contract the Si lattice by 0.014 Angstrom/atom% B added [139]. Therefore, heavy boron doping would compensate somewhat the strain in the SiGe layer. Based on the hardball sphere model, for complete strain compensation, the Ge to B concentration ratio is estimated to be ~6. However, the experimental results, based on the bow measurements on epitaxial wafers, give a higher ratio, ~8 [129]. In the case of E19/cm^3 boron doping, the amount of strain that boron atoms can compensate corresponds to a Ge concentration of 0.16 atom%. So the B compensation effect is not significant, since the average Ge concentration in the graded-Ge SiGe layer is 5–10 atom%.

3.6.3.2 Non-SEG Epitaxial SiGe Growth for HBT

The SiGe epitaxy for HBT is also carried out using non-SEG epitaxy. A method via batch processing was developed by IBM specifically for SiGe epitaxial for its HBT technologies, in which a blanket deposition is carried out at a low temperature (500°C–600°C) and at ultra high vacuum (UHV/CVD). Passivating the wafer surface with hydrogen and performing UHV/CVD at low temperatures, results in an epitaxial film that is practically defect free ($<10^3$ defects/cm^2) [141]. This low-temperature epitaxy (LTE) process enables abrupt, fully activated, in situ boron doping [142], and the controlled incorporation of germanium into the silicon lattice. It has been applied to bipolar and BiCMOS technologies, the LTE has replaced the implanted base (which is common in homojunction NPNs produced through conventional epitaxy) with an in situ grown base and graded germanium HBT [143]. The LTE process grows epitaxial silicon over the exposed HBT silicon regions of a patterned wafer, and polysilicon over regions that are protected by polysilicon or silicon dioxide. The SiGe epitaxial base region is contacted by the polysilicon that is formed over the shallow trench isolation (STI) during LTE deposition. In the HBT SiGe epitaxy, the boron doping profile is abrupt and with high concentration. To prevent boron out diffusion during processing, it is common to co-dope the layer with carbon by introducing SiH_3CH_3 during the epi deposition. It is believed that when the carbon atoms are incorporated onto substitutional sites, it can suppress boron out diffusion and maintain a more abrupt characteristic boron profile [144].

3.6.3.3 Selective Epitaxy for Elevated Source/Drain in CMOS

Since selective epitaxy was employed for depositing the base of a HBT for SiGeB and SiGeB:C, the next popular application of the selective epitaxy was for elevated S/Ds for CMOS devices. The raised S/D (RSD) is necessary where there is a limitation of silicon real estate. This is particularly true for SOI wafers, where the silicon layer thickness is limited. The "extension" of the S/D via silicon epitaxial deposition is to facilitate contact silicidation. The use of RSD is common and has been shown to enhance the technologies in 90 nm processing [145], ultra-thin SOI layer [146], as well as, for sSOI (strained silicon on insulator) [147]. An example of a S/D cross-section view in a transistor is shown in Figure 3.56. Unlike blanket epitaxial deposition on the starting substrates, the integration of a SEG process in a scaled CMOS technology needs to take into account the added thermal budget. The typical SEG thermal budget is high enough and could cause unwanted dopant diffusion (from a previously implemented process). Therefore, in the integration of the SEG in device processing, one needs to design the process and

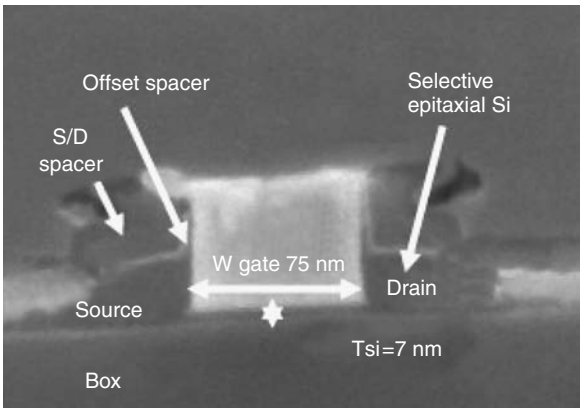


FIGURE 3.56 Scanning electron microscopy (SEM) image of an ultra thin SOI device with extended S/D by selective epitaxial deposition. (From Doris, B. et al., *IEDM Tech. Dig.*, 2003, 631. Reproduced with permission from IEEE.)

strategically place SEG in the processing sequence so as to avoid/minimize the adverse effects of the SEG thermal cycles [145].

3.6.3.4 Selective Epitaxy for Recessed S/D of PMOS

Since the use of strained silicon for a MOS transistor channel, in order to enhance carrier mobility, was demonstrated, several methods of fabricating strained silicon have been developed for CMOS device processing. The relevance of carrier mobility enhancement to the magnitude and sign of the silicon lattice strain is discussed in more detail in the literature [148,149,194,196,199]. As discussed, for maximum mobility enhancement, the *N*-channel silicon should be in tensile strain while the P channel favors compressive strain. Complementary metal–oxide–silicon fabrication on a strained silicon layer (such as in sSOI and sGOI wafers), where biaxial strain exists is a convenient way to realize the mobility enhancement, significantly in NMOS, although a smaller benefit for the PMOS (unless the SiGe substrate for strain layer growth has a Ge content of 35% or more [149]).

Another approach to implement a strain effect in CMOS channels was developed in which the localized uniaxial compressive strain in the *P*-channel MOSFET was induced by the deposited SiGe in a recessed S/D via SEG [150], while the tensile strain needed for the NMOS was induced by a nitride cap over the *N*-channel MOSFET. Figure 3.57 shows diagrams of such PMOS and NMOS transistors. A unique feature of the PMOS transistor is the embedding of a compressively strained SiGe film in the source and drain regions by using a SEG process. A combination of compressive SiGe strain and embedded SiGe S/D geometry induces a large uniaxial compressive strain in the channel region, thereby resulting in a significant hole mobility improvement. Greater than 50% strain induced hole channel mobility improvement is demonstrated for devices with 17% Ge composition [150]. It should be noted that a similar approach was proposed inducing a local tensile strain for the NMOS transistor by using silicon-doped with carbon instead of germanium [151]. These benefits are observed on long-channel MOSFETS, whereas scaled transistors with physical channel lengths at or below 45 nm or so typically exhibit only about 12%–15% improvement in mobility [197,198].

It is noted that the RSD SEG process can be combined with the recessed S/D in one epitaxial step and with different materials. Selective epitaxy will be important in FDSOI device processing where the silicon film is thin and in FinFET [152,196] fabrication, for wider fins. Recently, the SEG has also been employed in the Hybrid Orientation Technology (HOT) [191,192]. The fabrication is based on a SOI wafer, in which the handle wafer is of (110) surface orientation, while the top Si layer is of (100). Windows are opened on the SOI (100) layer (through the BOX and to expose handle wafer silicon), at the locations where (110) silicon is desired. Windows (with oxide liner as isolation) are then filled with (110) silicon

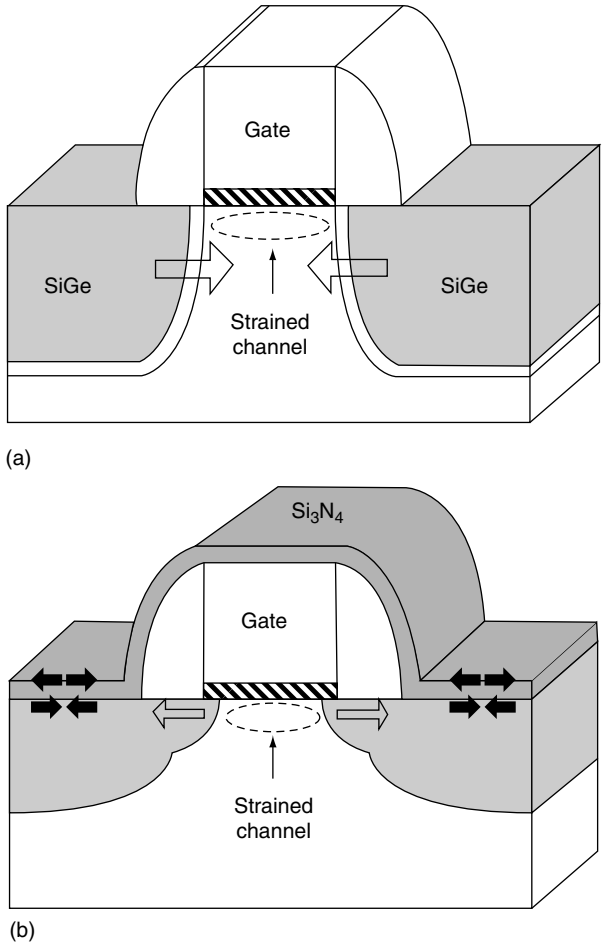


FIGURE 3.57 Models of transistors showing (a) compressive uni-axial strained in PMOS channel by recessed S/D selective epi, (b) tensile strain along NMOS channel by silicon nitride film on silicon. (From Liu, C., W., MaiKap, S., and Yu, C. Y., *IEEE Circuits Devices Mag.*, May/June, 2005. Reproduced with permission from IEEE.)

via SEG. Complementary metal–oxide–silicon processing on the finished planar wafer structure with PMOS and NMOS on (110) and (100) orientation surface regions, respectively, give the optimal performance in channel mobility, since PMOS exhibits highest mobility on (110) and NMOS exhibits highest mobility on (100) surfaces, when the channel is along a ⟨110⟩ direction. It is expected that certain device architectures will require multiple epitaxial steps integrated in its process. The trend in integration is to require SEG at even lower temperature process and improved selectivity. These demands present constant challenges for equipment improvement and deposition process development.

3.6.4 Strained Silicon Epitaxy

The introduction of elastic strain in the silicon transistor channel material is a widely accepted method for enhancing the carrier mobility of Si [153,193,195]. There are two general approaches in the strained Si technology that have been researched and demonstrated. The first method is to introduce local strain in the transistor channel region via transistor modular engineering during the transistor fabrication process. The strains induced during transistor processing are typically uniaxial (in one direction) and are

incorporated through tensile/compressive capping film layers or by recessed epitaxial film deposition in the S/D regions. The selective SiGe growth in the recessed S/D for a compressive strain in the PMOS channel was discussed above in the Epitaxial section. The second method is the introduction of global strain in the Si layer of the starting wafers via epitaxial Si deposition on a SeGe alloy substrate.

The bulk/global strained Si is analogous to today's Si CMOS p^-/p^+ substrate, which is composed of an epitaxial Si film grown on top of a uniform content SiGe alloy. The heteroepitaxy of silicon on a SiGe alloy is a convenient way to fabricate strained silicon with the resulting strain in the proper range to bring about the beneficial mobility enhancement effect in NMOS transistor fabrication. Since SiGe has a larger lattice constant than silicon, its lattice size increases with the Ge content. The thin silicon layer grown epitaxilly on SiGe will be in a biaxially tensile strained condition. However, the silicon layer thickness grown should not exceed the "critical thickness" criterion [132] to trigger strain relaxation via the formation of misfit dislocations. The degree of the strain so obtained in the silicon layer is proportional to the Ge content in the substrate and is customarily expressed in terms of it's Ge content. The practical and useful Ge content of the substrate is in the range of 15%–30% for a strain of approximately 1%. For today's CMOS fabrication, the strained Si technology is usually merged with SOI technology in two configurations: (1) sGOI (Strained Si on SiGe on Insulator)- strained Si film on top of the strain generating SiGe layer, with an insulating layer residing below it. (2) sSOI (Strained Si on Insulator)- merger of strained Si and SOI without the SiGe layer. The strained silicon layer on insulator is mechanically and thermally stable and is shown to maintain its strain up to 1100°C heat treatment temperature [154,155].

3.6.4.1 Growth of Relaxed SiGe-Substrate for Strained Silicon Epitaxy

Unlike the SiGe epitaxy discussed above for HBT applications which is in a pseudomorphic and strained condition, the SiGe layers prepared for the strained silicon growth have to be completely or highly relaxed and is in a state of equilibrium structure. Furthermore, as in all epitaxial growth, in order to minimize lattice defects in the strained silicon layer, the defect density in the SiGe substrate has to be minimized. The epitaxial growth of the SiGe substrate normally begins with a silicon substrate, generally of (100) orientation (for CMOS applications). Depending on the method used, the relaxed SiGe layer, with desired Ge concentration, may be obtained by layer growth with a gradual increase in Ge content (graded buffer approach) or via a shorter process of 2–3 steps, which may include depositions of SiGe layers with fixed Ge content, low temperature, Si cap layer, and high temperature anneal. Regardless of the approach used, the goal is to achieve a highly relaxed SiGe layer with the desired Ge content and minimum possible threading dislocation density ($\leq$E5/cm^2).

In the growth of SiGe layer on a silicon substrate with Ge in the concentration range of 10%–30%, one can easily exceed the critical thickness and cause the deposited layer to plastically relax with the generation of misfit dislocations at the interface. Ideally, one would prefer that all misfit dislocations are of pure edge dislocation type with Burgers vector $a/2$ [110], which is lying in the (001) plane, parallel to the epi-substrate interface. Since (001) is not a glide plane in a diamond cubic lattice, such dislocations are immobile except by the "climb" mechanism under stress. Therefore, these edge misfit dislocations would be "locked" at the interface, and one can achieve low defect density SiGe layer rather readily. However, in the SiGe/Si growth, many misfit dislocations are found to be of the 60° dislocation type. These dislocations and its threading components can glide on the inclined {111} planes and propagate during growth or under stress. The control of relaxation generated misfit dislocations from reaching the surface of the SiGe layer during epitaxial growth is the major consideration in deposition process design.

The graded buffer approach for growing relaxed SiGe layer is an established method [156–158], in which the SiGe deposition is carried out at a moderate temperature and the process begins at a low Ge content, 10% for example. The process gradually increases its Ge content to 20 or 30%, as the layer is deposited. The lattice mismatch strain is relaxed continuously via the generation of misfit dislocations during the growth process, until the desired Ge content is attained and the layer is fully relaxed. The bulk of the resulting misfit dislocations are contained in this buffer layer. At this time, the growth is continued for a few more microns to assure a fully relaxed structure, and to further "grow out" the propagating

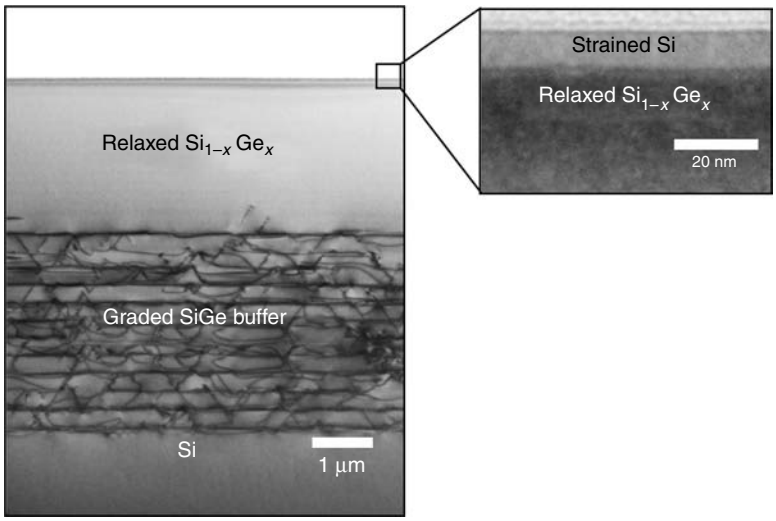


FIGURE 3.58 Transmission electron microscopy cross-sectional view of Si wafer showing SiGe substrate growth sequence via graded buffer approach. Insert shows strained Si layer deposited on the relaxed SiGe substrate. (Courtesy of Fitzgerald, E. A. Reproduced with permission.)

threading dislocations. The residual threading dislocation density attainable by the graded buffer approach is less than E5/cm^2. Figure 3.58 shows a TEM cross-section of the SiGe layer grown via the graded buffer approach and a 10 nm or so strained Si layer grown on top of the relaxed SiGe layer. Figure 3.59 is a lattice model describing configurations of atomic arrangement of epitaxial layers in a strained Si growing process via the graded SiGe buffer approach. The sequence in the model matches the growth sequence shown in Figure 3.58.

By its process nature, the graded buffer approach requires the growth of thicker SiGe buffer layers, 5–7 μm, in order to reach the desired Ge content, and at the same time, to establish relaxed structure with low defect density. Thicker layer and longer processing time can be a throughput factor in a manufacturing environment. Thicker SiGe film growth also results in an unfavorable surface morphology with a higher roughness [159].

A thinner SiGe layer deposition with discrete Ge content (non-graded) can be a shorter process, but the containment of the massive misfit dislocations and propagation of the threading components is the main challenge. Numerous approaches have been proposed toward this goal. One of the effective ways to control defects is to provide nucleation sites for misfit dislocation generation. These sites can be lattice defects intentionally induced in the silicon substrate or in the growing layer during the process. For example, low energy He implantation into the Si substrate was found to be effective in providing nucleation sites for misfit dislocation generation in the Si substrate, during the relaxation anneal of a deposited $Si_{0.7}Ge_{0.3}$/Si layer.[160]. The He bubbles form a narrow band of defects underneath the substrate/epi interface and contain the misfit defects upon annealing. Similarly, a thin layer of silicon deposited at a low temperature close to the amorphization temperature (LT–Si), is believed to generate saturated point defects, which can serve as nucleation centers for misfit dislocations generation during plastic relaxation of a strained SiGe layer. Using the LT–Si as nucleation sites, Chen et al. [161] demonstrated the defect containment effect of a low temperature deposited Si layer in $Si_{0.76}Ge_{0.24}$/Si deposition, resulting in a low threading dislocation density. This is shown in the TEM micrograph in Figure 3.60. Note that the dislocation loops generated are grown into the silicon substrate.

Utilizing the controlled defect generation of a low temperature Si cap (LT–Si), discussed above, a "two-step" process involving Si_xGe_{1-x} layer deposited at low temperature, followed by a high-temperature

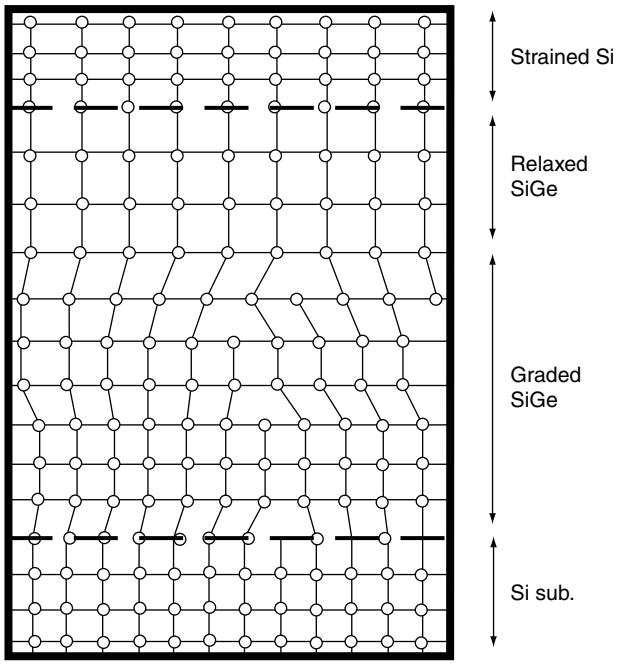


FIGURE 3.59 Lattice model showing transitions of atomic arrangements and state of lattice strain at each stage during the growth of strained Si via graded SiGe approach, starting from a Si substrate. The schematic matches the micrograph shown in Figure 3.59 in growing sequence. (From Liu, C. W., MaiKap, S., and Yu, C. Y., *IEEE Circuits Devices Mag.*, May/June, 2005. Reproduced with permission from IEEE.)

anneal, has been shown to result in a low threading dislocation density. Figure 3.61 shows a TEM micrograph of the cross-section of the $Si_{0.62}Ge_{0.38}$ layer grown via a two-step process [162]. In such a process, a low temperature GeSi layer with lower Ge content (~0.17%) was deposited after a LT–Si cap deposition on the Si substrate. This is followed by a high-temperature (~800°C) anneal to relax the strain with misfit dislocations nucleated spontaneously near the Si cap layer. This step results in a low defect

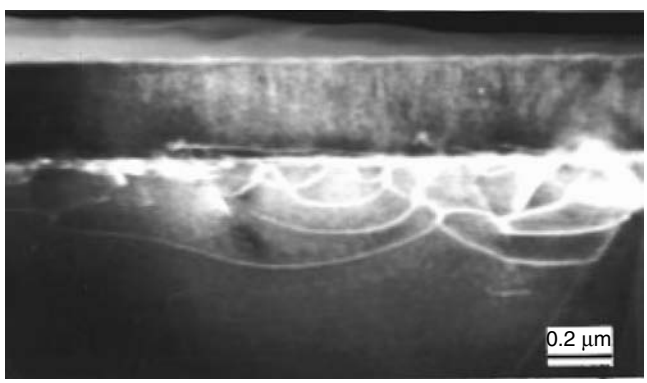


FIGURE 3.60 Transmission electron microscopy cross-sectional view of Si wafer after deposition of 50 nm LT–Si (400°C) and 300 nm Si 0.76 Ge 0.24 (550°C) followed by 800°C anneal for 20 min. The dislocation loops from the relaxation are observed to extend into the Si substrate, below the epi/substrate interface. (From Chen, H. et al., *J. Appl. Phys.*, 79, 1167, 1996. Reproduced with permission from AIP.)

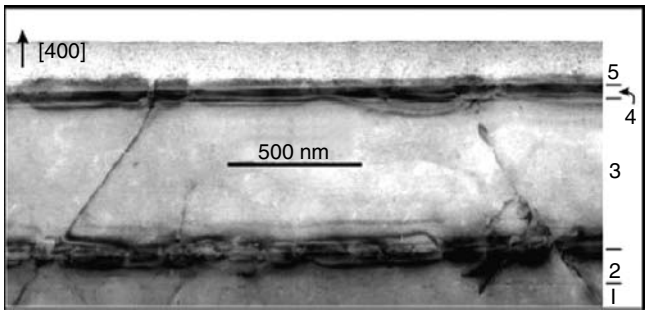


FIGURE 3.61 Transmission electron microscopy micrograph of Si (110) cross-section of a "two-step" process: 1 is the Si substrate, 2 are high-temperature and low-temperature Si buffers, 3 is the first $Si_{0.83}$ $Ge_{0.17}$ layer 4 is 50 nm low-temperature $Si_{0.83}$ $Ge_{0.17}$ layer grown right after the firstt layer was annealed and 5 is the second $Si_{0.62}$ $Ge_{0.38}$ layer. (From Bolkhovityanov, Y. B. et al., *Appl. Phys. Lett.*, 84, 4599, 2004. Reproduced with permission from AIP.)

density, $<$E6/cm^2 threading dislocations. Similarly, the procedures may be repeated to grow a relaxed GeSi layer with a higher Ge content, using the grown GeSi layer as a substrate. The process was shown to achieve a relaxed SiGe layer with Ge content up to 48%, with total layer deposition significantly below 1 μm (650 nm) and a density of threading dislocations close to E5/cm^2.

3.7 Oxygen Behavior in Silicon Processing

3.7.1 Oxygen Precipitation Kinetics

Oxygen in silicon tends to precipitate upon heat treatment when the oxygen concentration exceeds its solid solubility limit via homogeneous or heterogeneous nucleation mechanisms. The precipitation kinetics of oxygen at a given temperature can be described by a "time-temperature-transformation" (TTT) curve, such as that shown in Figure 3.62 for p^+ silicon, where the precipitation density is plotted as a function of annealing time at 1050°C. Although various models have been proposed to describe the basic nucleation/growth mechanism [163], no simple model can take into account the effect of all the factors involved, such as crystal thermal history, oxygen, dopant and carbon concentrations, point defect concentrations, and grown-in microdefects, as well as the detailed IC processes and sequence. Indeed, most of the understanding of the precipitation behavior has been based on experiments with mostly a few controlled parameters.

One of the most important factors affecting nucleation is the grownin microdefects. Although "dislocation-free" CZ silicon grown with today's technology is considered highly perfect, crystal lattice imperfections may be formed at, or behind, the freezing interface during growth. Microdefects, such as A-, B-, and C-type defects [84] provide sites for oxidation induced stacking faults at the wafer surface and oxygen precipitation in the bulk silicon upon subsequent heat treatment. Experiments have shown [165] that during a high-temperature anneal (without a previous low-temperature nucleation anneal), a heterogeneous nucleation mechanism dominates and microdefects play an important role in nucleation as the postcrystal growth "thermal history" experienced by the silicon. In this condition, precipitation in CZ silicon lacking microdefects would be retarded. On the other hand, when twostep anneal containing a low-temperature nucleation step initially is used, the homogeneous nucleation mechanism dominates, and the impact of grown-in microdefects on precipitation is diminished.

Oxygen precipitation kinetics at a given temperature is primarily a function of oxygen concentration. In the ASTM standard test methods [166] for the oxygen precipitation ability of *n*- and *p*-type CZ silicon, a high temperature (1050°C, 16 h, Method A) and a low-high (750°C, 4 h + 1050°C 16 h, Method B) thermal cycles are utilized. The combined A and B tests reveal the state of thermal history and

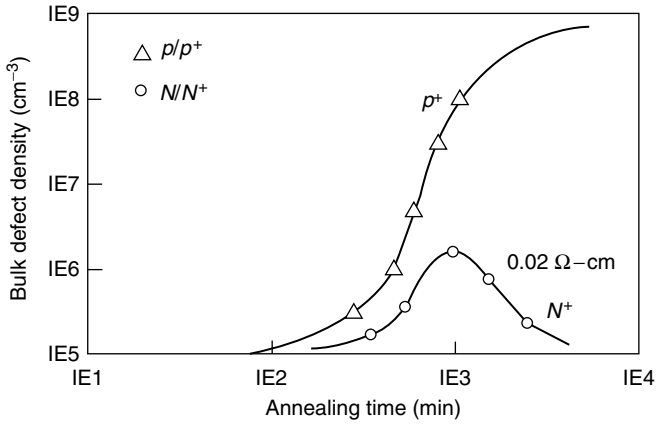


FIGURE 3.62 A comparison of oxygen precipitation kinetics in n^+ and p^+ annealed at 1050°C after epi deposition at 1150°C.

microdefect density of the wafers under test, when compared with a reference precipitation behavior. Figure 3.63 and Figure 3.64 show the oxygen precipitation as a function of oxygen concentration under test methods A and B, respectively, for wafers from several sources [167]. In this case, the wafers from different sources/suppliers have similar behavior.

Although it is well established that carbon in silicon has the benefit of oxygen precipitation, the precise control of carbon incorporation in a range of a few parts per million atomic during crystal growth has not been developed. While the control of carbon incorporation and uniformity are still issues of silicon growth, the carbon concentration in CZ crystals is usually kept to a minimum to avoid its influence.

The CVD deposited polysilicon film on the wafer backsurface was found to serve as a good extrinsic gettering sink because of its ability to generate dislocations and stacking faults at the backsurface. It was found that the backsurface deposited polysilicon also enhanced oxygen precipitation, which added to the

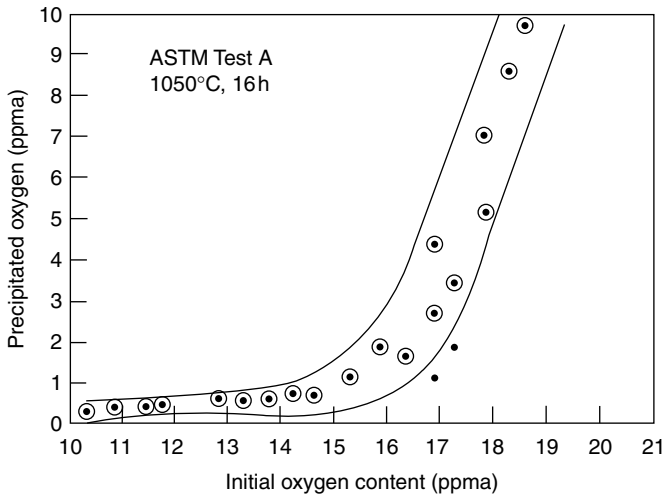


FIGURE 3.63 Oxygen precipitation after 1050°C, 16 h anneal. Data represent wafers from six sources. (From Swaroop, R. et al., *Solid State Technol.*, March, 1987, 85. Reproduced with permission from Solid State Technology.)

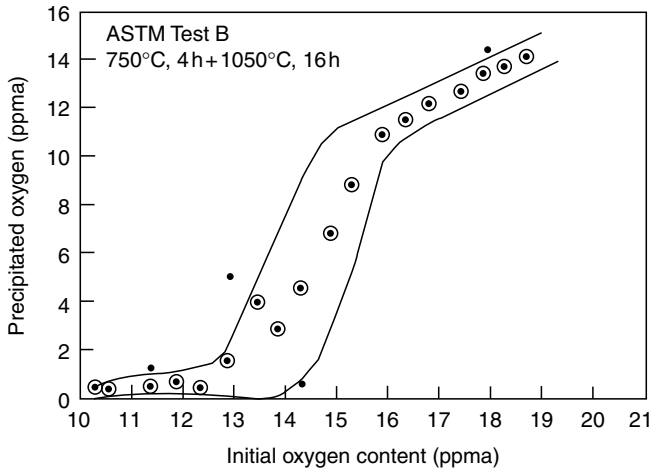


FIGURE 3.64 Oxygen precipitation after 750°C, 4 h + 1050°C, 16 h anneal. Data represent wafers from six sources. (From Swaroop, R. et al., *Solid State Technol.*, March, 1987, 85. Reproduced with permission from Solid State Technology.)

efficiency of "backside polysilicon gettering." The precipitation enhancement effect was thought to be a nucleation effect due to the polysilicon deposition thermal cycle. Shirai et al. [168] have shown that a polysilicon layer at the wafer backsurface causes localized oxygen precipitation near the backsurface. The enhancement effect was attributed to the absorption of silicon self-interstitials by the polysilicon, which causes an increase in vacancy concentration in the nearby silicon wafer, since [I][V] = Const. at equilibrium. Vacancies were considered to be effective nucleation sites.

3.7.2 Oxygen Precipitation in p^+ and n^+

In recent years, the use of n/n^+ and p/p^+ epitaxial structures in the VLSI/ULSI ICs has significantly increased, due to the structure's potential in reducing latch-up susceptibility in CMOS and alpha particleinduced soft error in memory devices. There is no distinguishable difference in oxygen precipitation kinetics between n- and p-type silicon in lightly doped silicon (typically, 10^{15} cm^{-3}). However, as the dopant is increased to the degenerately doped region of interested (0.005–0.02 ohm–cm), the behavior is drastically different depending on the conductivity type. In this range, oxygen precipitation is retarded in Sb-doped silicon while the kinetics are at their peak for ptype, boron doped silicon. Due to these properties, the p^+ substrate provides good internal gettering while Sb-doped n^+ leads to reduced microdefect formation and intrinsic gettering during silicon processing. Figure 3.62 compares oxygen precipitation kinetics in p^+ and n^+ wafers under isothermal anneal at 1050°C, after an epitaxial deposition at 1150°C (without pre-epi heat treatment). The possible sources of differences in the precipitation kinetics in p^+ and n^+ silicon include oxygen incorporation, oxygen diffusivity, grown-in microdefect density, and nucleation mechanism.

It is now established that oxygen incorporation into silicon crystals depends on the type and concentration of the dopant [43]. On the average, p^+ crystals exhibit about 25% higher incorporation rate than n^+ crystals (with p^+ contents higher than p, while n^+ are lower). The difference in oxygen incorporation cannot account for the difference observed in precipitation kinetics in p^+ and n^+. The diffusion mechanism of oxygen in n^+ was found to be the same as in p^- silicon [42]. The difference between the precipitation behavior in p^+ and n^+ appears to reside in the nucleation mechanism. Various models of precipitation mechanisms have been proposed in order to explain the kinetics in p^+ and n^+.

Early models [169,170] were based on the assumption that Si self-interstitials are responsible for heterogeneous nucleation; complexing of Sb^+I^- would reduce the nucleation centers in n^+. Later models favored nucleation mechanisms based on vacancies as nucleation centers, VO complexes [41,47,171,172]. Suppression of precipitation in n^+ was suggested as due to reduction of V^- by the formation of a high concentration of V^-Sb^+ complex [47,171]. The enhancement effect in p^+ may be due to the formation of B^-I^+ complexes and increasing V^- [47], since IV = Const.

3.7.3 Effects of Oxygen Precipitation on Device Processing

Oxygen precipitation is generally considered beneficial for its impurity gettering function in silicon processing. However, uncontrolled or untimely precipitation can result in adverse effects. In the following, the ill-effects of an uncontrolled oxygen precipitation are discussed in terms of defect generation, both for epitaxial and bulk wafers.

Fast precipitation kinetics of oxygen in p^+ silicon is difficult to control. One such consideration is during pre-epitaxial deposition thermal treatments of p^+ substrates for film growth, or deposition for wafer backseal (to minimize autodoping), or external gettering purposes. For example, 2-h, 800°C thermal treatment can induce a significant precipitation in the case of a high oxygen, e.g., 19 ppma (ASTM F121-80), p^+ wafer without the formation of a precipitate-free region (denuded zone; DZ). Such fast precipitation can generate a configuration where precipitates and bulk stacking faults would intersect the polished wafer surface. Subsequent silicon epitaxial growth on such defected surfaces will generate a unique crystallographic defect structure appearing as pairs of etch pits, or "viper pits," on the p/p^+ epitaxial wafer surface, after preferential etching [173]. The pairs of pits correspond to the two dislocations, initiated from a single strain center, an oxygen precipitate (or related dislocation loops) at the epitaxial–substrate interface.

One can avoid unwanted oxygen precipitation in the pre-epi thermal processing by using low-oxygen p^+ silicon and/or shorter, lower temperature thermal cycles. However, it is desirable that precipitation occurs during the epitaxial deposition. Some precipitation during the epitaxial deposition can add gettering ability and reduce nucleation centers for stacking faults, and saucer pit density in the epitaxial layer. For the same reason, oxygen precipitations are desired in the critical thermal steps following epitaxial deposition in the IC processing, such as gate oxidation. Therefore, in an ideal situation, the p^+ substrate shall be tailored such that (a) there is little or no precipitation as a result of pre-epi thermal treatment, (b) some precipitation occurs during epitaxial deposition to provide an internal gettering and (c) full precipitation occurs early in the device fabrication steps. This is illustrated in Figure 3.65. In p/p^+ wafers for conventional CMOS applications, the formation of a (denuded zone; DZ) in p^+ substrate beneath the epi layer is not a critical feature. In principle, no harm can result even DZ = 0, so long as it is formed after epi deposition. However, in some ULSI designs with trench structures fabricated on thin p/p^+ epi, the large-aspect-ratio trench structures usually extend beyond the epi layer and into the p^+ substrate. In these applications, a DZ is required so that trenches can be fabricated in the defect-free p^+ region. For device processing based on lightly doped silicon, n- or p-type, oxygen precipitation kinetics are much slower than in p^+. Depending on the oxygen level and thermal cycle sequence, the precipitate distribution in the wafer cross-section can be classified into four categories as shown schematically in Figure 3.66 the configuration in Figure 3.66(a) is ideal, where significant precipitation occurs along with a sufficient defect-free region for device fabrication. This situation can occur during the IC processing thermal cycles or it may be brought about by added pre-IC-processing thermal cycles, such as Hi–Lo–Hi or Hi–Lo thermal treatment schemes. In the latter case, "full precipitation" usually occurs as a result of preheat treatments at the early stage of device processing.

In the cases where precipitation gettering is important, steps for enhanced precipitation are often used. However, care must be taken to avoid over-precipitation, which would "drain" interstitial oxygen from the silicon crystals. The reduction of interstitial oxygen resulting in a significant oxygen precipitation lowers the wafer yield strength and makes it more susceptible to plastic deformation when subjected to thermal gradients. The net results are wafer warpage and the generation of slip dislocations into active

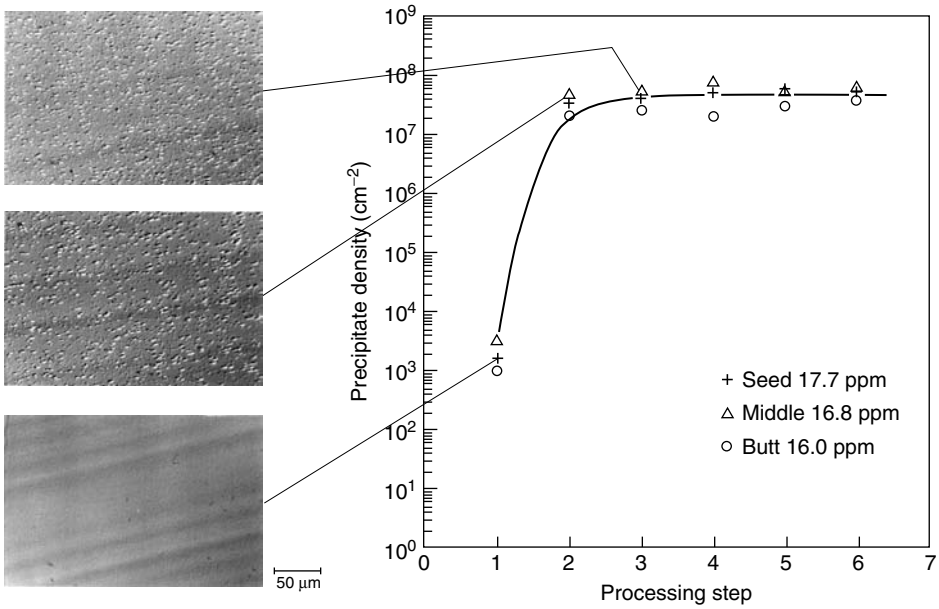


FIGURE 3.65 Oxygen precipitation density in p^+ silicon after thermal cycles due to backseal (step 1) epi deposition (step 2) and first four steps of Twin-Tub V process technology. (From Lin, W., *Semiconductor Silicon 1990*, eds. Huff, H. R., Barraclough, K. G., and Chikawa, J.-i, Electrochemical Society, Pennington, NJ, 1990, 569. Reproduced with permission from Electrochemical Society.)

device regions. It has been shown [174] that at least 10 ppma of interstitial oxygen should be retained to prevent the onset of slip dislocations during processing of a bipolar device. The configuration in Figure 3.66(b) is the result from non-perfect DZ formation, due to local precipitation enhancement caused by grown-in microdefects or micro-oxygen concentration fluctuations. This results in oxygen precipitations/bulk stacking faults near the active region of the wafer surface layer, causing emitter-collector shorts, source-drain shorts, transistor junction leakage, and DRAM storage or capacitance related yield problems. The configuration in Figure 3.66(c) corresponds to a uniform oxygen precipitation where oxygen out-diffusion is impeded, resulting in no DZ formation and the possibility that precipitates/bulk stacking faults would intersect the wafer surface. The situation stems from a gross oxygen concentration-process, thermal cycle-mismatch, or the use of a dominant Lo–Hi-type process cycle at the beginning of the thermal processing. When oxygen precipitates intersect the wafer surface, it can cause oxide defects and poor breakdown characteristics. The reliability effect of thin gate oxides grown on a silicon surface containing a high density of oxygen precipitates and bulk stacking faults induced by heat treatment has been investigated [175]. Both voltage-ramp and time dependent dielectric breakdown (TDDB) results indicate that oxygen precipitates cause significant gate oxide degradation. Figure 3.51 shows the defect density (D_0) extracted from the voltage measurements, illustrating that the "heat treated" wafers have larger D_0 than "normal" bulk wafer and p/p^+ epitaxy for $V>5$ V.

Figure 3.66(d) represents a situation where the wafer's oxygen content is below the threshold of oxygen precipitation for the thermal cycles used and no precipitation is formed. However, some precipitation can be triggered near the wafer back surface for gettering by pre-processing wafer backside polysilicon film deposition via mechanism discussed above. In the era of deep sub-micrometer processing technology using large diameter-wafers, there has been a trend to use lower oxygen concentrations [176] for "soft internal gettering," with targeted bulk precipitation density to be about 5×10^5/cm^3 and DZ width of approximately 15 μm. The main advantage of using a low oxygen silicon is the decrease in

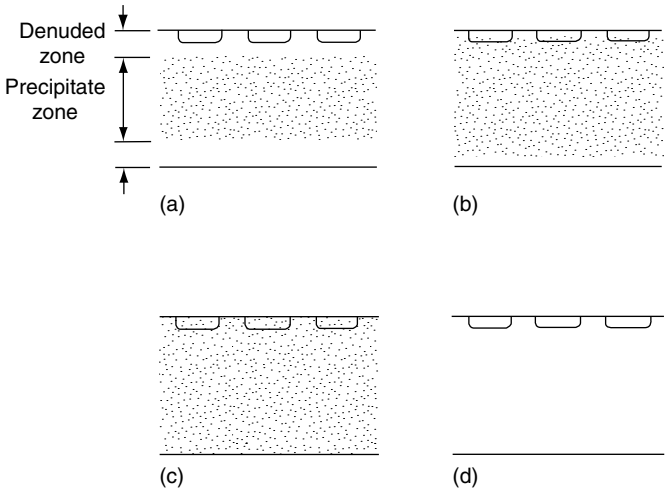


FIGURE 3.66 Schematics of wafer cross-sections showing four categories of DZ/precipitation distributions during integrated circuit (IC) process. (From Lin, W., *Semiconductor Silicon 1990*, eds. Huff, H. R., Barraclough, K. G., and Chikawa, J.-i, Electrochemical Society, Pennington, NJ, 1990, 569. Reproduced with permission from Electrochemical Society.)

the possibility of microdefect formation near the wafer's active region. Microdefects are critical in memory devices. Furthermore, the possibility of wafer warpage and slip dislocation generation induced by over-precipitation of oxygen is also reduced. It is believed that the trend for lower oxygen is likely to continue in view of the advancements made in IC contamination control in IC fabrication lines; class one or better environments, cleaner process equipment, and improvements in the purity of process, chemicals, and materials. The potential for heavy metal contamination is accordingly reduced and so is the IC's dependence on internal gettering by oxygen precipitates. When additional gettering capability is needed, an external gettering method may be added, such as backside polysilicon deposition.

The foregoing discussions and examples on silicon processing of both epitaxial and bulk wafers illustrate the importance of oxygen control in silicon, whether or not precipitation gettering is wanted. Uncontrolled precipitation can result in precipitate-induced defects during epitaxial growth, electrically sensitive defects in the device, active regions, and inferior gate oxides. In these situations, the harmful effects of oxygen can outweigh its beneficial aspects. In applications where internal gettering is desired, it is essential to "engineer" oxygen precipitation to occur at the optimal locations and at the right time through tight control of wafer oxygen concentration and/or strategically tailoring of the thermal processes. Alternatively, one can positively avoid the ill-effects of oxygen precipitates by utilizing silicon with a sufficiently low oxygen concentration for "soft" or no precipitation. In either case, the control of oxygen with a narrow concentration distribution and minimization of grown-in microdefects in silicon are necessary.

3.7.4 Denuded Zone and Oxygen Precipitation by Controlled Vacancy Concentration

In recent years, a different approach has been proposed [178] to optimize the DZ and IG configuration, based on a point defect engineering scheme, in which the precipitate density and its depth distribution are decoupled from the interstitial oxygen concentration and thermal history of the crystal growth. By utilizing the significant differences in diffusivities between interstitial, vacancy and oxygen ($I > V \gg O$) [179,180], it is possible to separate point defects in close proximity to the surface, using high-

temperature rapid thermal processing (RTP). The RTP heating and cooling effects erase the oxygen precipitation nuclei formed during the crystal growth, and create a unique vacancy profile (actually an out-diffused profile) that would retard the oxygen precipitation near the wafer free surface, while enhancing the precipitation in the wafers bulk in the subsequent anneal. Effectively, in this scheme, the oxygen precipitate concentration is controlled by the vacancy concentration created, not by the oxygen concentration or the grown-in defect history. High-temperature RTP heating generates a high concentration of vacancy-interstitial pairs. They are then separated by differential diffusion to the wafer free surface, leaving a dominant vacancy concentration in the wafer bulk. During the cool down, the vacancies try to maintain equilibrium concentration by further diffusion to the surface, leaving a vacancy-depleted near-surface region, 50–100 μm in depth. Due to the low vacancy concentration, which is below the critical value for initiating oxygen clustering, a precipitation-free DZ (referred to as "Magic Denuded Zone" (MDZ) [178] is produced during the subsequent low temperature anneal. Meanwhile, high precipitate density is generated in the high-density vacancy bulk region for IG.

It is important to realize that the vacancy controlled precipitation scheme via RTP is affected by the processing ambient. This is demonstrated in Figure 3.67. The vacancy profiles in Figure 3.66(a) were determined by a Pt diffusion technique [180] for samples prepared in separate argon and nitrogen RTP annealing ambients. The etched wafer cross-sections in (b) and (c) show two very different precipitation densities at the wafer's near surface region, resulting from two different vacancy profiles under argon (MDZ) and nitrogen annealing ambients, respectively, followed by a Lo-Hi precipitation anneal [178]. The RTP anneal under argon allows vacancy out diffusion, and vacancy depletion in the wafer near surface, as discussed above for MDZ. On the other hand, the RTP annealing in nitrogen results in Si interstitial undersaturation (due to the formation of silicon nitride at the surface) in the near surface region causing vacancy injection in order to maintain equilibrium. Therefore, depending on the RTP heating time, cooling rate and ambient gas used, one can arrive at a range of different vacancy profiles from which the oxygen precipitate distributions are determined. Thus, the wafer defect engineering via vacancy concentration profile control by RTP certainly opens a new avenue to effect internal gettering by oxygen precipitates.

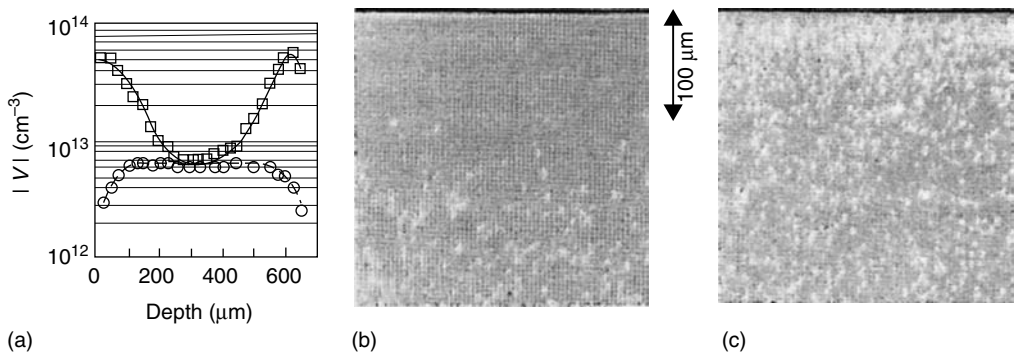


FIGURE 3.67 (a) Vacancy concentration obtained using Pt diffusion Jacob, M. et al., Determination of vacancy concentrations in the bulk of silicon wafers by platinum diffusion experiments, *J. Appl. Phys.*, 82, 182, 1997 on two wafer samples after 1250°C RTP for 30 s in nitrogen (squares) or argon (circles). After subsequent Lo–Hi precipitation treatment (4 h @ 800°C + 16 h @ 1000°C), the etched wafer cross-sections micrographs are shown in (b) argon annealed (Magic denuded zone; MDZ) and (c) nitrogen annealed. (From Rozgonyi, G. A., *Semiconductor Silicon 2002*, eds. Huff, H. R., Fabry, L., and Kishino, S., Electrochemical Society, Pennington, NJ, 2002, 149. Reproduced with permission from Electrochemical Society.)

3.8 Other and New Applications of Silicon Materials

3.8.1 Nitrogen Doping and Its Effects on CZ Silicon

Nitrogen atoms are electrically neutral and are located at interstitial sites in the silicon lattice when its concentration is below the solubility limit [181]. Nitrogen doping of CZ silicon, in the concentration of 10^{13}–10^{15} atoms/cm^3 has been shown to have several significant effects on silicon properties and its applications during thermal processing. Among them are strengthening effect, oxygen precipitation enhancement, and reduction/retardation of void defects (D defects). Nitrogen impurity was demonstrated in FZ wafers [182] to increase slip resistance and to reduce warpage during thermal processing in CZ wafers [183,184]. These mechanical strengthening effect is due to the dislocation pinning effects of the nitrogen atoms in the lattice. Effectively, nitrogen atoms immobilize dislocations although nitrogen atoms dispersed in the crystal lattice have no appreciable effect on the velocity of dislocations in motion [185]. The immobilization is related to the gettering of nitrogen atoms by the dislocation core. Oxygen atoms in silicon lattice have the same effect. However, the strength of the locking per one nitrogen atom is about 30 times greater than that of oxygen atom [185]. Thus nitrogen atoms enhance the yield strength of a silicon crystal when the crystal is initially dislocated.

In recent years, nitrogen doping during the CZ crystal growth (N-CZ) at a level comparable to the vacancy concentration, ~E14–E15/cm^3, have shown the effect of creating N–V or N–V–O complexes to enhance the nucleation of oxygen precipitates and to simultaneously reduce the size and density of voids (D-defects), that is, COPs. It is believed that in the normal CZ crystal growth, during cool down, void formation takes place near ~1070°C, before the formation of grown-in oxygen precipitate nuclei (~1040°C). In the presence of nitrogen atoms, stable complexes containing nitrogen atoms form at a higher temperature with excess vacancy concentration, before the void formation (at a lower temperature). Two consequences can result: one is that the formation of the nitrogen containing complexes consume available excess vacancies, and therefore, reduce the density as well as the size of the void defects [112]. The nitrogen doping causes COPs to change their morphology from octahedral to smaller triclinic platelet or be annihilated [186]. Secondly, the N–V and N–V–O complexes formed at high temperatures are more stable and in higher density than "grown-in" oxygen precipitate nuclei without nitrogen doping. The oxygen precipitation enhancement effect increases with increasing nitrogen doping level (see Figure 3.68). The precipitation enhancement effect is significant in low-thermal budget device processing where internal gettering is desired, but involving low oxygen wafers. With the reduced size of D defects (COPs) in the grown CZ when doped with nitrogen, it also facilitates the high temperature COP annihilation, annealing operation (1200°C) to reduce the COP size and density (in the subsurface region). Figure 3.69 shows the effect of nitrogen content on annealing efficiency in 300 mm wafers.

3.8.2 High Resistivity Silicon

The silicon materials used in today's CMOS and BICMOS fabrications are mostly lightly boron doped bulk substrates, 5–25 ohm–cm, or p^-/p^+ epitaxial substrates. As the CMOS technology continues to scale down the design rule, allowing the circuits to operate at the gigahertz frequency range, it provides an opportunity for the low cost integration of RF/digital/analog functions on the same chip for system on chip (SOC) applications. For such integration, the high resistivity silicon substrate is favored. Unlike digital CMOS, RF circuitry requires low noise. High resistivity substrate reduces capacitively coupled crosstalk between digital, analog and RF components. The high resistivity also reduces substrate "losses" (in form of eddy current) [188] and improves the quality factor of the inductor. "High resistivity" normally refers to resistivity $\geq$1000 ohm–cm, although in practice, substrates with resistivity of 50 ohm–cm or higher are used for improvements in performance.

As discussed in Section 3.2 of this chapter, silicon single crystals grown from high purity polysilicon (99.999999999% pure) via FZ method can achieve much higher resistivity than CZ method. This is

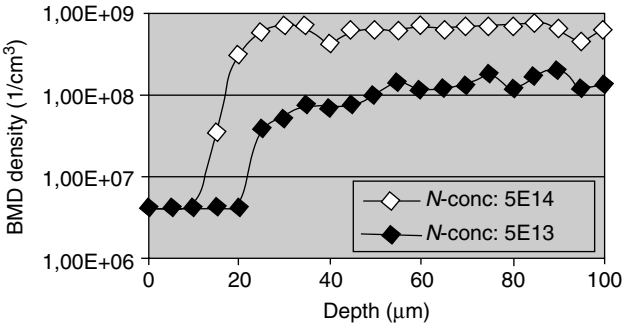


FIGURE 3.68 Bulk micro defect (BMD) density and denuded zone (DZ) measurement by scanning infrared microscope on annealed 300 mm wafer (1200°C/2 h) for two different nitrogen concentrations. (From Muller, T. et al., *Semiconductor Silicon 2002*, eds. Huff, H. R., Fabry, L., and Kishino, S., Electrochemical Society, Pennington, NJ, 2002, 194. Reproduced with permission from Electrochemical Society.)

mainly due to its melt which does not contact any substances other than the ambient inert gas during growth. The large volume of FZ consumed today is in the resistivity range of 10–200 ohm–cm. Higher resistivity silicon can be manufactured readily. In fact, FZ crystals with very high resistivity from 9000 to 30,000 ohm–cm can be grown using multiple-pass vacuum process. The CZ silicon is usually prepared to resistivity of 25 ohm–cm or less due to contaminations from the fused silica (SiO_2) crucibles used in the process.

The major contaminant in CZ silicon is oxygen, ~E18 atoms/cm^3. This level of interstitial oxygen content provides the strengthening effect and enables the function of internal gettering by oxygen precipitation in the CZ wafers during IC processing. These properties have been considered important attributes of the substrate for thermal processing. While being very low in oxygen content, ~E16/cm^3, or approximately 0.2 ppma, the FZ wafers are lacking these beneficial effects in thermal processing. For these reasons, CZ wafers have been used almost exclusively for IC fabrications for the last four decades. Therefore, for high-resistivity silicon for advanced IC fabrications, CZ silicon is still preferable to FZ.

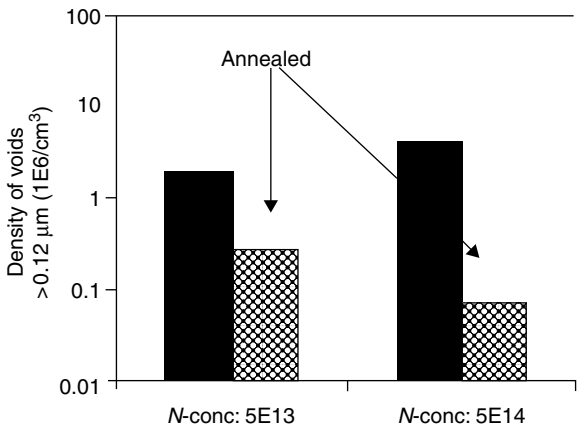


FIGURE 3.69 Comparison of void density on low and medium nitrogen doped 300 mm silicon before and after anneal (1200°C/1 h). (From Muller, T. et al., *Semiconductor Silicon 2002*, eds. Huff, H. R., Fabry, L., and Kishino, S., Electrochemical Society, Pennington, NJ, 2002, 194. Reproduced with permission from Electrochemical Society.)

There are two major issues in growing and preparing high resistivity CZ silicon. First, in the CZ silicon growth, high resistivity corresponds to a low intended doping level. The contaminants (such as Group III and V elements) dissolved from the silica crucible will have greater impact on the doping accuracy and the high-resistivity limit. Boron contamination is a concern, which is an impurity in the quartz sand, used for fused silica crucible fabrication. Complete removal of boron impurity from fused silica is not an easy task. One can avoid the contaminants from the natural quartz sand by using synthetic SiO_2 for fabrication of the silica crucible and to achieve higher resistivity. But significantly higher costs will prevent using this approach for high-volume production. However, with the use of high purity silica crucible, or crucible coated with high purity material, CZ silicon with 1000 ohm–cm or higher may be obtained.

The second issue is that the high resistivity CZ silicon is susceptible to more interference from the "low temperature oxygen donor" formed at 300°C–500°C (highest formation rate is at 450°C), arising from the formation of oxygen–silicon complexes, SiO_x. Oxygen donors are formed in as-grown crystals during crystal cool down following crystal growth. They can be annihilated by annealing the wafers at 650°C–700°C, followed by quenching in an inert ambient. The "true" resistivity of the grown silicon can then be measured. However, donors can be formed again during the wafer processing when experiencing temperatures in the neighborhood of 450°C and the remaining [Oi] is significant. In CMOS processing, considerable "back end" processing occurs near 450°C, and can provide an environment for thermal donor regeneration. Thermal donor formation causes resistivity change and dopant compensation in the substrate (even conductivity type can change). In some cases, degradation of device parametric performance has been attributed to donor formation. Since thermal donor concentration is a power function (power >1) of interstitial oxygen concentration [189], maintaining low value of [Oi] is essential during the "back end" processing. Two general approaches may be used: (1) to incorporate a Hi–Lo–Hi-type thermal cycle before or at the beginning of an IC processing sequence, in order to form a DZ near the wafer surface (oxygen is depleted in the active region) and a highly precipitated bulk (oxygen is consumed in precipitates). Care must be taken not to over precipitate oxygen which causes softening of the wafer strength [174]; (2) to grow high-resistivity silicon using MCZ method for low [Oi] such that only a few thermal donors may be formed during IC processing [190]. An MCZ wafer with low [Oi] can still provides a mechanical strengthening effect; however, low or no oxygen precipitation is to be expected for the benefit of internal gettering. See configuration in Figure 3.66(d).

3.9 Summary

Modern Ultra Large Scale Integration ICs, fabricated with design rules approaching 45 nm, depends on the availability of highly perfect silicon single crystals, which are prepared exclusively by the CZ crystal growth method. The silicon material prepared today, polished wafer or epitaxial wafer, must meet the challenges imposed by the stringent requirements of the starting materials for current and future design rule generations, as guided by the Starting Materials Road Map of the ITRS.

In parallel with the design rule decrease, both chip size and circuit complexity increase. There is a continued push for an increased wafer diameter in order to reduce IC manufacturing cost. In the up-scaling of the wafer diameter, in the 300–450 mm range, the most significant technical challenges are in the crystal growth. The major issues include crystal weight support, reduced growth rate, and growing long crystals from "ultra-large" charge capacity and/continuous growth. The dislocation-free as-grown yield will dominate the cost of manufacturing 300–450 mm-diameter wafers. Furthermore, the grown-in microdefects, have vital impacts on the GOI and proper function of memory devices. With in-depth understanding of the relationship between intrinsic point defect concentration and crystal growing parameters, the microdefect-free crystal growth processes (i.e., for crystals free of D and A defects) have been developed. To circumvent the microdefect problem in polished wafers, p^-/p^+ epi and hydrogen/argon-annealed wafers continue to be widely used for IC fabrication. MCZ is the growth method for large-capacity, large diameter silicon crystals growth for reducing thermal convection and better control

of oxygen. Consequently, the interstitial oxygen contents of the large diameter silicon ($\geq$300 mm) are generally lower than the smaller diameter silicon grown with normal CZ. Lower oxygen content makes it difficult, if not impossible, to exercise internal gettering by oxygen precipitation under the low thermal budget of deep sub micron design rule device processing.

Heteroepitaxy technology has played an important role in the manufacturing of emerging strained silicon as starting materials and in fabricating device modules via selective epitaxy during IC processing, such as HBT base, RSD, recessed S/D, and FinFET. These epitaxial depositions cause either enhancement in device performance or solve Si real estate limitations.

As the design rule reduces the requirements for wafer mechanical dimension control, particulate, as well as metal contamination, become more stringent. To achieve improved flatness for sub-100 nm IC fabrication, new techniques in wafer shaping and polishing have been implemented, these include surface grinding in conjunction with reduced chemical etching, double-side polishing, and plasma assisted chemical etching. Finally, for improved wafer contamination control, purer starting raw materials, process chemicals, proper equipment maintenance, and wafer handling are essential.

Semiconductor silicon materials have been the backbone for IC fabrication since its infancy. As IC technology evolved through several generations, the silicon materials have also gone through many changes in order to meet the challenges of design rule reduction and demands for improvement in device performance. Today's silicon materials are results of decades' continued effort in research and development. While this review covers science and engineering of silicon materials preparation and its behavior upon processing, the history of silicon materials research and development for the last 50 years, and current applications of silicon materials in the IC are subjects of recent review papers by Huff [200] and Tsuya [201], respectively.

References

1. Pfann, W. G. "Principles of Zone-Melting." *Trans. Am. Inst. Min. Metall. Eng.* 194 (1952): 747.
2. Theuerer, H. C. U.S. Patent 3,060,123, 1962.
3. Keller, K., and A. Muhlbauer. *Floating Zone Silicon*. New York: Marcel Dekker, 1981.
4. Meese, J. M., ed. *Neutron Transmutation Doping in Semiconductors*. New York: Plenum, 1979.
5. Bullis, W. M. "Oxygen Concentration Measurements." In *Oxygen in Silicon*, edited by F. Shimura, New York: Academic Press, 1994 chap. 4.
6. Leroy, B., and C. Plougonven. "Warpage of Silicon Wafers." *J. Electrochem. Soc.* 127 (1980): 961.
7. Takasu, S. et al. "Wafer Bow and Warpage." *Jpn. J. Appl. Phys.* 20, no. Suppl. 1 (1981): 25.
8. Patel, J. R., and A. R. Chaudhuri. "Oxygen Precipitation Effect on the Deformation of Dislocation-Free Silicon." *J. Appl. Phys.* 33 (1963): 2223.
9. Yonenaga, I., K. Sumino, and K. Hoshi. "Mechanical Strength of Silicon Crystals as a Function of Oxygen Concentration." *J. Appl. Phys.* 56 (1984): 2346.
10. Sumino, K., I. Yonenaga, and A. Yusa. "Mechanical Strength of Oxygen-Doped Float-Zone Silicon Crystals." *Jpn. J. Appl. Phys.* 19 (1980): L763.
11. Abe, T. et al. "Impurities in Silicon Single Crystals—A Current View." In *Semicouductor Silicon 1981*, edited by H. R. Huff, R. J. Kriegler, Y. Takeishi et al., 5. Princeton, NJ: Electrochemical Society, 1981.
12. Jastrzebski, L. et al. "The Effect of Nitrogen on the Mechanical Properties of Float Zone Silicon and on CCD Device Performance." *J. Electrochem. Soc.* 134 (1987): 466.
13. Czochralski, J. "Ein Neues Verfahren Zur Messung der Kristallisationsgeschwindigkeit der Metalle." *Z. Phys. Chem.* 92 (1918): 219.
14. Teal, G. K., and J. B. Little. "Growth of Germanium Single Crystals." *Phys. Rev.* 78 (1950): 647.
15. Abe, T. "Crystal Fabrication." In *VLSI Electronics Microstructure Science*, Vol. 12, edited by N. G. Einspruch, and H. R. Huff, 3. New York: Academic Press, 1985.
16. Dash, W. C. "Silicon Crystals Free of Dislocations." *J. Appl. Phys.* 29 (1958): 736.

17. Lin, W., and D. W. Hill. "Large-Diameter Czochralski Silicon Crystal Growth." In *Silicon Processing ASTM STP 804*, 24. Philadelphia, PA: ASTM, 1983.
18. Lin, W., and K. E. Benson. "The Science and Engineering of Large Diameter Czochralski Silicon Crystal Growth." *Ann. Rev. Mater. Sci.* 17 (1987): 273.
19. Billig, E. *Proc. R. Soc. Lond. Ser.* A235 (1956): 37.
20. Van Vleck, L. H. *Elements of Materials Science and Engineering*. 3rd ed. Reading, MA: Addison-Wesley, 1975, chap. 8.
21. Jordan, A. S., R. Caruso, and A. R. Von Neida. *Bell Syst. Tech. J.* 59 (1980): 593.
22. Pfann, W. G. *Zone Melting*. 2nd ed. New York: Wiley, 1966 chap. 2.
23. Benson, K. E., W. Lin, and E. P. Martin. "Fundamental Aspects of Czochralski Silicon Crystal Growth for VLSI." In *Semiconductor Silicon 1981*, edited by H. R. Huff, R. J. Kregler, and Y. Takeishi, 33. Princeton, NJ: Electrochemical Society, 1981.
24. Fiegl, G. "Recent Advances and Future Directions in CZ-Silicon Crystal Growth Technology." *Solid State Technol.* August (1983): 121.
25. Lorenzini, R. E., A. Iwata, and K. Lorenz. U.S. Patent 4,036,595, 1977.
26. Craven, R. A. et al. "The Effect of Carbon on Czochralski Silicon Used for Dynamic Random Access Memory Production." *Proc. Mat. Res. Soc. Symp.* 59 (1986): 359.
27. Trumbore, F. A. "Solid Solubility of Impurity Elements in Germanium and Silicon." *Bell Syst. Tech. J.* 39 (1960): 205.
28. Barraclough, K. G., and P. Ward. *J. Electrochem. Soc. Meet. Ext. Abstr.* 167 (1983): 474.
29. Gilmore, D. et al. "The Impact of Furnace Graphite Parts on CZ-Grown Single Crystal Silicon's Radial Impurity Distribution." In *High Purity Silicon IV*, edited by C. L. Claeys et al., 102. Pennington, NJ: Electrochemical Society, 1996.
30. Kodera, H. *Jpn. J. Appl. Phys.* 2 (1963): 212.
31. Burton, J. A., R. C. Prim, and W. P. Slichter. "The Distribution of Solute in Crystal Grown from the Melt." *J. Chem. Phys.* 21 (1987): 1953.
32. Cochran, W. G. *Proc. Cambridge Phil. Soc.* 30 (1934): 365.
33. Carruthers, J. R., A. F. Witt, and R. E. Reusser. "Czochralski Growth of Large Siliconcrystal—Convection and Segregation." In *Semiconductor Silicon 1977*, edited by H. R. Huff, and E. Sirtl, 61. Princeton, NJ: Electrochemical Society, 1977.
34. Carruthers, J. R. *J. Electrochem. Soc.* 114 (1967): 959.
35. Carlberg, T., T. B. King, and A. F. Witt. *J. Electrochem. Soc.* 129 (1982): 189.
36. Hoshikawa, K. et al. "Control of Oxygen Concentration in CZ Silicon Growth." In *Semiconductor Silicon 1981*, edited by H. R. Huff, R. J. Kregler, and Y. Takeishi, 101. Princeton, NJ: Electrochemical Society, 1981.
37. Carruthers, J. R., and K. Nassau. "Nonmixing Cells Due to Crucible Rotation During Czochralski Crystal Growth." *J. Appl. Phys.* 39 (1968): 5205.
38. Kakimoto, K. et al. *J. Cryst. Growth* 88 (1988): 356.
39. Kakimoto, K. et al. *J. Cryst. Growth* 89 (1989): 412.
40. Moody, J. W. "Oxygen in Czochralski Crystals and Melts—A Review." In *Semiconductor Silicon 1986*, edited by H. R. Huff, T. Abe, and B. Kolbeson, 100. Pennington, NJ: Electrochemical Society, 1986.
41. Tsuya, H., Y. Kondo, and M. Kauamori. *Jpn. J. Appl. Phys.* 22 (1983): L16.
42. Oates, A. S., and W. Lin. "Temperature Dependence of Interstitial Oxygen Diffusion in Antimony-Doped Czochralski Silicon." *Appl. Phys. Lett.* 53 (1988): 2659.
43. Oates, A. S., and W. Lin. "Infrared Measurements of Interstitial Oxygen in Heavily Doped Silicon." *J. Cryst. Growth* 89 (1988): 117.
44. Walitzki, H. et al. "Control of Oxygen and Precipitation Behavior of Heavily Doped Silicon Substrate Materials." In *Semiconductor Silicon 1986*, edited by H. R. Huff, T. Abe, and B. Kolbesen, 86. Pennington, NJ: Electrochemical Society, 1986.
45. Nozaki, T. et al. "Behavior of Oxygen in the Crystal Formation and Heat Treatment of Silicon Heavily Doped with Antimony." *J. Appl. Phys.* 59 (1986): 2562.
46. Itoh, Y. In *Proceedings of 31st Applied Physics Conference*, Kawasaki, Japan, 609, 1984.

47. Shimura, F. et al. In *VLSI Science and Technology 1985*, edited by M. W. Bullies, and S. Broydo, 507. Pennington, NJ: Electrochemical Society, 1985.
48. Barraclough, K. G., and R. W. Series. In *Reduced Temperature Processing for VLSI*, edited by R. Reif, 452. Pennington, NJ: Electrochemical Society, 1986.
49. Darken, L. S., and R. W. Gurry. *Physical Metallurgy of Metals*, 287. New York: McGraw Hill, 1953.
50. Kaiser, W., P. H. Keck, and C. F. Lange. "Infrared Absorption and Oxygen Content in Silicon and Germanium." *Phys. Rev.* 101 (1956): 264.
51. Yatsurugi, T. et al. *J. Electrochem. Soc.* 120 (1973): 975.
52. Lin, W., and D. W. Hill. "Oxygen Segregation in Czochralski Silicon Growth." *J. Appl. Phys.* 54 (1983): 1082.
53. Lin, W., and M. Stavola. "Oxygen Segregation and Microscopic Inhomogeneity in Czochralski Silicon." *J. Electrochem. Soc.* 132 (1985): 1412.
54. Harada, H. et al. In *VLSI Science and Technology 1985*, edited by M. W. Bullies, and S. Broydo, 526. Pennington, NJ: Electrochemical Society, 1985.
55. Iino, E. et al. "Formation of Interstitial Oxygen Striations in CZ Silicon Single Crystals." In *Semiconductor Silicon 1994*, edited by H. R. Huff, W. Bergholz, and K. Sumino, 148. Pennington, NJ: Electrochemical Society, 1994.
56. Lin, W. "Oxygen Segregation and Microscopic Inhomogeneity in CZ Silicon," In *Proceedings of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 288, 1996.
57. Shimura, F. et al. *Semiconductor Silicon Technology*, 258. New York: Academic Press, 1989.
58. Yamagishi, H. et al. "Evaluation of FPDs and COPs in Silicon Single Crystals." In *Semiconductor Silicon 1994*, edited by H. R. Huff, W. Bergholz, and K. Sumino, 124. Pennington, NJ: Electrochemical Society, 1994.
59. Park, J. G. et al. In *The Physics and Chemistry of SiO_2 and the Si–SiO_2 Interface II*, edited by C. R. Helms, and B. E. Deals, 168. New York: Plenum, 1993.
60. Nakajima, K. et al. "Distribution of As-Grown Defects in a Silicon Single Crystal." In *Semiconductor Silicon 1994*, edited by H. R. Huff, W. Bergholz, and K. Sumino, 168. Pennington, NJ: Electrochemical Society, 1994.
61. Nishikawa, H. et al. In *Proceedings of the 45th Symposium on Semiconductor and Integrated Circuits Technology*, Japanese Electrochemical Society, Japan, 100, 1993.
62. Marioton, B. P. R., and U. Gosele. "Transport of Thermodynamic Information by Self-Interstitials between Precipitates in Silicon." *J. Appl. Phys.* 63 (1988): 4661.
63. Patrick, W. J., S. J. Scilla, and W. A. Westdorp. U.S. Patent 4,010,064, 1977.
64. Secco d'Aragona, F. U.S. Patent 4,545,849, 1985.
65. Lin, W., and C. W. Pearce. "Properties of Uniform Oxygen Czochralski Silicon Crystals." *J. Appl. Phys.* 51 (1980): 5540.
66. Watanabe, M. et al. "Oxygen-Free Silicon Crystal Grown from Silicon Nitride Crucible." In *Semiconductor Silicon 1981*, edited by H. R. Huff, R. J. Kriegler, and T. Takeishi, 126. Pennington, NJ: Electrochemical Society, 1981.
67. Watanabe, M. *ASTM Symposium on Semiconductor Processing*, Abstract, San Jose, 1984.
68. Utech, H. P., and M. C. Flemings. "Elimination of Solute Banding in Indium Antimonide Crystals by Growth in Magnetic Field." *J. Appl. Phys.* 37 (1966): 2021.
69. Witt, A. F., C. J. Herman, and H. C. Gatos. *J. Mater. Sci.* 5 (1970): 822.
70. Chandrasekhar, S. *Philos. Mag.* 43, no. 7 (1952): 501.
71. Hoshi, K. et al. *Electrochem. Soc. Meet. Ext. Abstr.* 157 (1980): 811.
72. Takasu, S. et al. "Si Crystal Growth Under Magnetic Field." In *Semiconductor Silicon 1990*, edited by H. R. Huff, K. G. Barraclough, and J.-i. Chikawa, 45. Pennington, NJ: Electrochemical Society, 1990.
73. Series, R. W. et al. *Electrochem. Soc. Meet Ext. Abstr.* 167 (1985): 396.
74. Braggins, T. T., and R. N. Thomas. *Electrochem. Soc. Meet. Ext. Abstr.* 169 (1986): 351.
75. Hoshi, K. et al. *J. Electrochem. Soc.* 132 (1985): 693.

76. Ohwa, M. et al. "Growth of Large Diameter Silicon Single Crystal under Horizontal or Vertical Magnetic Field." In *Semiconductor Silicon 1986*, edited by H. R. Huff, T. Abe, and B. Kolbesen, 117. Pennington, NJ: Electrochemical Society, 1986.
77. Hoshi, K., N. Isawa, and T. Suzuki. *ULSI* August (1986): 51.
78. Series, K. G. et al. *Electrochem. Soc. Meet. Ext. Abstr.* 167 (1985): 396.
79. Braggins, T. T., and R. N. Thomas. *Electrochem. Soc. Meet. Ext. Abstr.* 169 (1986): 354.
80. Hirata, H., and K. Hoshikawa. *J. Cryst. Growth* 96 (1989): 47.
81. Hirata, H., and K. Hoshikawa. *J. Cryst. Growth* 98 (1989): 777.
82. Series, R. W. *J. Cryst. Growth* 97 (1989): 92.
83. Hoshikawa, K. and H. Hirata. "Control of Oxygen Concentration in Czochralski Silicon Crystal Growth by a CUSP Magnetic Field." In *Proceedings of 2nd International Symposium of Advanced Science and Technology of Silicon Mateials*, Kono, 85, 1996.
84. de Kock, A. J. R. "Point Defect Condensation In Dislocation-Free Silicon Crystals." In *Semiconductor Silicon 1977*, edited by H. R. Huff, and E. Sirtl, 508. Princeton, NJ: Electrochemical Society, 1977.
85. Marcus, R. B. In *VLSI Technology*, edited by S. M. Sze, 543. New York: McGrow Hill, 1983.
86. de Kock, A. J. R., and W. M. van de Wijgert. *J. Cryst. Growth* 49 (1980): 719.
87. Ryuta, J. et al. *Jpn. J. Appl. Phys.* 29 (1990): L1947.
88. Yamagishi, H. et al. *Semicond. Sci. Technol.* 7 (1992): A135.
89. Gall, P. et al. In *Defect Control in Semiconductors*, edited by K. Sumino, 255, North Holland, Amsterdam, 1990.
90. Umeno, S. et al. *Jpn. J. Appl. Phys.* 36 (1997): L591.
91. von Ammon, W. et al. *J. Cryst. Growth* 151 (1995): 273.
92. Dornberger, E., and W. von Ammon. *J. Electrochem. Soc.* 143 (1996): 1648.
93. Voronkov, V. V. *J. Cryst. Growth* 59 (1982): 625.
94. Voronkov, V. V., R. Falster, and J. C. Holzer. In *Crystalline Defects and Contamination: Their Impact and Control in Device Manufacturing II*, edited by B. O. Kolbeson et al., 3. Pennington, NJ: Electrochemical Society, 1997.
95. von Ammon, W. "Crystal Growth of Large Diameter CZ Si Crystal." In *Proceeding of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 233, 1996.
96. Suhren, M. et al. "Crystal Defects in Highly Boron Doped Silicon." In *High Purity Silicion IV*, edited by C. L. Claeys et al., 132. Pennington, NJ: Electrochemical Society, 1996.
97. Wagner, P. et al. "Surface and Crystal Defects of Czochralski Silicon Wafers, their Relations and Modification by Chemical and Thermal Processes." In *Proceeding of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 101, 1996.
98. Wijaranakula, W., and S. Archer. *J. Electrochem. Soc.* 143 (1996): 1636.
99. Park, J. G. et al. "Structure and Morphology of d-Defects in CZ Si." In *Semiconductor Silicon 1994*, edited by H. R. Huff, W. Bergholz, and K. Sumino, 370. Pennington, NJ: Electrochemical Society, 1994.
100. Izumi, M. et al. "Segregation Behavior During Silicon Single Crystal Growth." *J. Appl. Phys.* 78 (1995): 5984.
101. Ueki, N., M. Itsumi, and T. Takeda. In *International Conference on Solid State Devices and Materials*, Yokohama, LA-1, 862, 1996.
102. Itsumi, M. et al. "Structure and Nature of Octahedral Void Defects Observed in Standard CZ–Si." In *Proceeding of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 270, 1996.
103. Takeno, H., M. Kato, and Y. Kitagawara. "Morphology and Nature of Grown-in Microdefects in Czochralski Silicon Crystals." In *Proceeding of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 294, 1996.
104. Dornberger, D. et al. "The Impact of Dwell Time above 900°C During Crystal Growth on the Gate Oxide Integrity of Silicon Wafers." In *High Purity Silicon IV*, edited by C. L. Claeys et al., 140. Pennington, NJ: Electrochemical Society, 1996.
105. Hourai, M. et al. *J. Electrochem. Soc.* 142 (1995): 3193.

106. Tanahashi, K., N. Inoue, and Y. Mizokawa. "Formation of Polyhedron Voids in CZ Silicon." In *Proceeding of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 51, 1996.
107. Kim, K. M., and P. Smetana. *J. Cryst. Growth* 100 (1990): 527.
108. Yamagishi, H. et al. "CZ Crystal Growth Development in Super Silicon Crystal Project." In *Proceeding of 2nd International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 59, 1996.
109. Kuramoto, M. "Super Silicon Initiative and Future Large Wafer Size Diameter." In *Semiconductor Silicon 2002*, edited by H. R. Huff, L. Fabry, and S. Kishino, 163. Pennington, NJ: Electrochemical Society, 2002.
110. Chiou, H. D., T. Y. T. Lee, and S. Teng. *J. Electrochem. Soc.* 144 (1997): 2881.
111. Furuya, H., K. Harada, and J.-G. Park. "Defect Reduction and Improved CZ Single—Crystal Silicon." *Solid State Technol.* June (2001): 109.
112. Rozgonyi, G. A. "Control of Point Defects, Impurities and Extended Defects in CZ Si: The Original/Ongoing Silicon Nanoscale Engineering Defect Science." In *Semiconductor Silicon 2002*, edited by H. R. Huff, L. Fabry, and S. Kishino, 149. Pennington, NJ: Electrochemical Society, 2002.
113. Kubota, H. et al. "Perfect Silicon Surface by Hydrogen-Annealing." In *Semiconductor Silicon 1994*, edited by H. R. Huff, W. Bergholz, and K. Sumino, 225. Pennington, NJ: Electrochemical Society, 1994.
114. Shimizu, Y. et al. *Jpn. J. Appl. Phys.* 36 (1997): 2565.
115. Arai, Y. et al. "The Growth of Silicon Single Crystals by the Magnetic Field Applied Continuous CZ (CMCZ) Method." In *Semiconductor Silicon 1994*, edited by H. R. Huff, W. Bergholz, and K. Sumino, 180. Pennington, NJ: Electrochemical Society, 1994.
116. Shiraishi, Y., S. Kurosaka, and M. Imai. *J. Cryst. Growth* 166 (1996): 685.
117. SEMI International Standards *Materials Volume*. (1994). Mountain View, 15. CA: SEMI International, 1994.
118. Robbins, H., and B. Schwartz. *J. Electrochem. Soc.* 107 (1960): 108.
119. Moreland, J. A. "The Technology of Crystal and Slice Shaping." In *VLSI Electronic Microstructure Science*. Vol. 12, edited by N. G. Einspruch and H. R. Huff, 63, Academic Press: New York, 1985.
120. Bollinger, D. and C. B. Zarowin. In *Advances in Fabrication and Metrology for Optical and Large Optics*, Vol. 966, 82, Soc. of Photo-Optical Instrumentation Engineers: Bellingham, WA, 1988.
121. Kern, W., and D. A. Puotinen. "Cleaning Solution Based on Hydrogen Peroxide for Use in Silicon Semiconductor Technology." *RCA Rev.* 31 (1970): 187.
122. Kern, W. "Purifying Si and SiO_2 Surfaces with Hydrogen Peroxide." *Semicond. Int.* April (1984): 94.
123. Sun, R. C., and J. T. Clemens. "Characterization of Reverse-Bias Leakage Current and Their Effect on the Holding Time Characteristics of MOS Dynamic RAM Circuit." *IEDM Tech. Dig.* (1977): 254.
124. Clemens, J. T. et al. U.S. Patent No. 4216489, 1980.
125. Chatterjee, P. K. et al. "Leakage Studies in High-Density Dynamic MOS Memory Devices." *IEEE Trans. Electron. Devices* ED-26 (1976): 564.
126. Aoki, M., T. Itakura, and N. Sasaki. "Gettering of Iron Impurities in p/p+Silicon Wafers with Heavily Bonor-Doped Substrates." *Appl. Phys. Lett.* 66 (1995): 2709.
127. Benton, J. L. et al. "Iron Getter Mechanism in Silicon." *J. Appl. Phys.* 80 (1996): 3275.
128. Horn, F. H. "Densitometric and Electrical Investigation of Boron in Silicon." *Phys. Rev.* 97 (1955): 1521.
129. Lin, W. et al. "Misfit Stress and Dislocations in p/p+ Epitaxial Silicon Wafers; Effect and Elimination." In *Defect in Silicon II*, edited by M. Bullis, 161. Pennington, NJ: Electrochemical Society, 1991.
130. Washburn, J., G. Thomas, and J. H. Queisser. "Diffusion Induced Dislocations." *J. Appl. Phys.* 35 (1964): 1909.
131. Weertman, J., and J. R. Weertman. *Elementary Dislocation Theory*, 54. New York: MacMillan, 1965.
132. Matthews, J. W., and A. E. Blakeslee. "Defect in Epitaxial Multilayers I. Misfit Dislocations." *J. Cryst. Growth* 27 (1974): 118.

133. Burghartz, J. N. et al. "Selective Epitaxy Base Transistor." *IEEE Elec. Dev. Lett.* 9 (1988): 299.
134. Endo, N. et al. "Novel Device Isolation Technology with Selective Epitaxial Growth." *IEDM Tech. Dig.* (1982): 242.
135. Stivers, A. R., C. H. Ting, and J. O. Borland. In *Chemical Vapor Deposition 1987*, edited by G. W. Cullen, 389. Pennington, NJ: Electrochemical Society, 1987.
136. Pai, C. S. et al. "Chemical Vapor Deposition of Selective Epitaxial Silicon Layer." *J. Electrochem. Soc.* 137 (1990): 971.
137. Borland, J. O., and C. I. Drowley. "Advanced Dielectric Isolation through Selective Epitaxial Growth Techniques." *Solid State Technol.* 28 (1985): 141.
138. Olesinski, W., and G. J. Abbaschian. *Bull. Alloy Phase Diag.* 5 (1984): 180.
139. Horn, F. H. "Densitometer and Electrical Investigation of Boron in Silicon." *Phys. Rev.* 97 (1955): 1524.
140. Meyerson, B. S. "Silicon: Germanium-Based Mixed-Signal Technology for Optimization of Wired and Wireless Telecommunications." *IBM J. Res. Dev.* 44, no. 3, (2000).
141. Meyerson, B. S. "Low Temperature Silicon Epitaxy by Ultrahigh Vacuum/Chemical Vapor Deposition." *Appl. Phys. Lett.* 48 (1986): 797.
142. Meyerson, B. S. et al. "Non Equilibrium Boron Effects in Low-Temperature Epitaxial Silicon Film." *Appl. Phys. Lett.* 50 (1987): 113.
143. Harame, D. L. et al. "Epitaxial Base Transistor with Ultrahigh Vacuum Chemical Vapor Deposition (UHV/CVD) Epitaxy: Enhanced Profile Control for Greater Flexibility in Device Design." *IEEE Electron Device Lett.* 10 (1989): 156.
144. Rucker, H. et al. "Dopant Diffusion in C-doped Si and SiGe: Physical Model and Experimental Verification." *IEDM Tech. Dig.* (1999): 345.
145. Park, H. et al. "High Performance CMOS Devices on SOI for 90 nm Technology Enhanced by RSD (Raised Source/Drain) and Thermal Cycle//Spacer Engineering." *IEDM Tech. Dig.* (2003): 635.
146. Doris, B. et al. "Device Design Considerations for Ultra-Thin SOI MOSFETs." *IEDM Tech. Dig.* (2003): 631.
147. Rim, K. et al. "Fabrication and Mobility Characteristics of Ultra-Thin Strained Si directly on Insulator (SSDOI) MOSFETs." *IEDM Tech. Dig.* (2003): 47.
148. Takagi, S. et al. "Comparative Study of Phonon-Limited Mobility of Two-Dimensional Electrons in Strained and Unstrained Si on Metal-Oxide-Semiconductor Field Effect Transistors." *J. Appl. Phys.* 80 (1996): 1567.
149. Oberhuber, R., G. Zandler, and P. Vogel. "Subband Structure and Mobility of Two-Dimensional Holes in Strained Si/SiGe MOSFETs." *Phys. Rev.* B58 (1998): 9941.
150. Ghani, T. et al. "A 90 nm High Volume Manufacturing Logic Technology Featuring Novel 45 nm Gate Length Strained Silicon CMOS Transistors." *IDEM Tech. Dig.* (2003): 987.
151. Murthy, A. et al. US Patent No. 6,621,131. September 16, 2003.
152. Kedzierski, J. et al. "Metal Gate FinFET and Fully-Depleted SOI Devices Using Total Gate Silicidation." *IEDM Tech. Dig.* (2002): 247.
153. Liu, C. W., S. MaiKap, and C. Y. Yu. "Mobility-Enhancement Technologies." *IEEE Circuits Devices Mag.* May/June, (2005).
154. Mazure, C. "The Smart Cut Technology for Future IC Technology." In *Proceedings of 4th International Symposium on Advanced Science and Technology of Silicon Materials*, Kona, HI, 281, November, 2004.
155. Ghyselen, B. et al. "Strained Silicon on Insulator Wafers Made by the Smart Cut™ Technology." *Mat. Res. Soc. Symp. Proc.* 809 (2004): B2.3.1.
156. Fitzgerald, E. A. et al. "Totally Relaxed Ge_xSi_{1-x} Layers with Low Threading Dislocation Density Grown on Si Substrate." *J. Appl. Lett.* 59 (1991): 811.
157. Samavedam, S. B., and E. A. Fitzgerald. "Novel Dislocation Structure and Surface Morphology." *J. Appl. Phys.* 81 (1997): 3108.
158. Herzog, H. J. et al. "SiGe-Based FETs: Buffer Issues and Device Results." *Thin Solid Films* 380 (2000): 36.

159. Fitzgerald, E. A., and S. B. Samavedam. "Point and Surface Defect Morphology of Graded, Relaxed GeSi Alloys on Si Substrates." *Thin Solid Films* 294 (1997): 3.
160. Luysberg, M. et al. "Effect of Helium Implantation and Annealing on the Relaxation Behavior of Pseudomorphic Si_{1-x} Ge_x Buffer Layers on Si (100) Substrates." *J. Appl. Phys.* 92 (2002): 290.
161. Chen, H. et al. "Low Temperature Buffer for Growth of a Low-Dislocation-Density SiGe Layer on Si by Molecular Beam Epitaxy." *J. Appl. Phys.* 79 (1996): 1167.
162. Bolkhovityanov, Y. B. et al. "Enhanced Strain Relaxation in a Two-Step Process of Ge_xSi_{1-x}/Si(001) Heterostructures Grown by Low-Temperature Molecular Beam Epitaxy." *Appl. Phys. Lett.* 84 (2004): 4599.
163. See for example,Wada, K., and N. Inoue. In *Semiconductor Silicon 1986*, edited by H. R. Huff, T. Abe, and B. Kolbesen, 778. Pennington, NJ: Electrochemical Society, 1986.
164. Petroff, P. M., and A. J. R. de Kock. *J. Cryst. Growth* 30 (1975): 117.
165. Lin, W., and A. S. Oates. "Anomalous Oxygen Precipitation in Czchralski Silicon." *Appl. Phys. Lett.* 56 (1990): 128.
166. Annual Book of ASTM Standards, Vol. 10.05, 1993.
167. Swaroop, R. et al. "ASTM Oxygen Precipitation Study." *Solid State Technol.* March (1987): 85.
168. Shirai, H., A. Yamaguchi, and F. Shimura. "Effect of Back-Surface Polycrystalline Silicon Layer on Oxygen Precipitation in Czochralski Silicon Wafers." *Appl. Phys. Lett.* 54 (1989): 1748.
169. de Kock, A. R., and W. M. van de Wijgert. *J. Cryst. Growth* 49 (1980): 718.
170. de Kock, A. R., and W. M. van de Wijgert. "The Influence of Thermal Point Defects on the Precipitation of Oxygen in Dislocation-Free Silicon Crystals." *Appl. Phys. Lett.* 38 (1981): 888.
171. Wijaranakula, W., J. Matlock, and H. Mollenkopf. "The Effect of Pre- and Post-Epitaxial Annealing on Oxygen Precipitation and Internal Gettering in N/N++(100) Epitaxial Wafers." In *Semiconductor Fabrication*, edited by D. Gupta *ASTM STP 990*, 371. Philadelphia, PA: ASTM, 1989.
172. Pearce, C. W. *Mat. Res. Soc. Symp. Proc.* 36 (1985): 231.
173. Lin, W. "Oxygen Behavior in Silicon Processing." In *Semiconductor Silicon 1990*, edited by H. R. Huff, K. G. Barraclough, and J.-i. Chikawa, 569. Pennington, NJ: Electrochemical Society, 1990.
174. Lin, W., and K. G. Moerschel. "Oxygen Precipitation and Intrinsic Gettering in Bipolar Integrated Circuit Processing." In *Reduced Temperature Processing for VLSI*, edited by R. Reif, 438. Pennington, NJ: Electrochemical Society, 1986.
175. Boyko, K. C., R. L. Feiller, and W. Lin. 1990. Unpublished.
176. Watanabe, M. *Solid State Technol.* April (1991): 133.
177. Lin, W. "The Incorporation of Oxygen into Silicon." In *Oxygen in Silicon*, edited by F. Shimura, New York: Academic Press, 1994, chap. 2.
178. Falster, R. et al. "The Engineering of Si Wafer Materials Properties through Vacancy Concentration Profile Control." In *High Purity Silicon 1998 PV 98-13*, 135. Pennington, NJ: Electrochemical Society, 1998.
179. Pagani, M. et al. "Spatial Variations in Oxygen Precipitation in Silicon after High Temperature Rapid Thermal Annealing." *Appl. Phys. Lett.* 70 (1997): 1573.
180. Jacob, M. et al. "Determination of Vacancy Concentrations in the Bulk of Silicon Wafers by Platinum Diffusion Experiments." *J. Appl. Phys.* 82 (1997): 182.
181. Yatsurugi, T. et al. *J. Electrochem. Soc.* 120 (1973): 975.
182. Abe, T. et al. "Impurities in Silicon Single Crystals." In *Semiconductor Silicon 1981*, edited by H. R. Huff, 54. Pennington, NJ: Electrochemical Society, 1981.
183. Abe, T. et al. In *Symposium on VLSI Science and Technology*, edited by W. M. Bullis, 543. Pennington, NJ: Electochemical Society, 1985.
184. Chiou, H. D. et al. In *Symposium on VLSI Science and Technology*, edited by K. Bean, 59. Pennington, NJ: Electrochemical Society, 1984.
185. Sumino, K. et al. "Effect of Nitrogen on Dislocation Behavior and Mechanical Strength in Silicon Crystals." *J. Appl. Phys.* 54 (1983): 5016.
186. Nakai, K. et al. "Formation of Grown-In Defects in Nitrogen Doped CZ–Si Crystals." In *Proceedings of 4th International Symposium on Advanced Science and Technology of Silicon Materials*, Kono, 88, 2000.

187. Muller, T. et al. "Argon Annealed 300 mm Wafers Complementing pp- Epitaxial Layers." In *Semiconductor Silicon 2002*, edited by H. R. Huff, L. Fabry, and S. Kishino, 194. Pennington, NJ: Electrochemical Society, 2002.
188. Frei, M. R. et al. "Integration of High-Q Inductors in a Latch-Up Resistant CMOS Technology." *IEDM Tech. Dig.* (1999): 757.
189. Wagner, P. "Infrared Absorption Studies of Thermal Donors in Silicon." In *Proceedings of MRS Symposium on Oxygen, Carbon, Hydrogen and Nitrogen in Crystalline Silicon*, Vol. 59, edited by J. C. Mikkelsen, et al., 125, Boston, 1985.
190. Ohguro, T. "Performance of Digital-Analog Mixed Device on a Si Substrate with Resistivity beyond 1 k ohm–cm." *IEDM Tech. Dig.* (2000): 757.
191. Yang, M. et al. "High Performance CMOS Fabricated on Hybrid Substrate with Different Crystal Orientations." *IEDM Tech. Dig.* (2003): 453.
192. Doris, A. "A Simplified Hybrid Orientation Technology (SHOT) for High Performance CMOS." *Symp. VLSI Technol.* (2004): 86.
193. Lee, M. L. et al. "Strained Silicon, SiGe and Ge Channels for High Mobility Metal-Oxide-Semiconductor Field Effect Transistors." *J. Appl. Phys.* 97 (2005): 011101.
194. Huff, H. R., and P. M. Zeitzoff. "A Perspective of Strained Silicon." *Solid State Technol.* January, (2004).
195. Huff, H. R., and P. M. Zeitzoff. "A Perspective on Enhancing Mobilities." *Solid State Technol.* January (2004): 26.
196. Huff, H. R., and P. M. Zeitzoff. "An Analytical Look at Vertical Transistor Structures." *Solid State Technol.* January (2004): 59.
197. Wang, H. C. et al. "Substrate-Strained Silicon Technology: Process Integration." *IEDM Tech. Dig.* (2003): 61.
198. Sanuki, T. et al. "Scalability of Strained Silicon CMOSFET and High Drive Current Enhancement in 40 nm Gate Length Technology." *IEDM Tech. Dig.* (2003): 65.
199. Huff, H. R. et al. "Starting Materials and Functional Layers for the 2005 International Technology Roadmaps for Semiconductors: Challenges and Opportunities." In *Characterization and Metrology for ULSI Technology 2005, AIP Conference Proceedings*, 788, 39.
200. Huff, H. R. "An Electronics Division Retrospective (1952–2002) and Future Opportunities in the Twenty First Century." *J. Electrochem. Soc.* 149 (2002): S35–S58.
201. Tsuya, H. "Present Status and Prospect of Si Wafers for Ultra Large Scale Integration." *Jpn. J. Appl. Phys.* 43, no. 7A (2004): 4055.

4

SOI Materials and Devices

4.1 Introduction **4**-2

4.2 SOI Basics: A Tutorial **4**-3
What It Is: A Definition • How It Is Made: A Very Brief Description • How SOI Devices Work

4.3 SOI Wafer Fabrication Methods-Some Details **4**-6
Wafer Bonding • Smart Cut™ Technology • Bond and Etchback • ELTRAN • SIMOX • Other Fabrication Methods

4.4 Advanced Wafer Engineering **4**-18
Crystal Orientations • Strained Silicon-on-Insulator • Ge-on-Insulator and Other on-Insulator Substrates

4.5 Physical Characterization of SOI Wafers **4**-24
Si and BOX Thickness Measurements • Surface Roughness • Structural Defects • Stress (Strain) Measurements by Raman Spectroscopy • Inspection for Particles and Other Defects

4.6 Electrical Characterization **4**-26
Wafer Characterization: Ψ-MOSFET • Device Characterization

4.7 Partially-Depleted SOI MOSFETs **4**-28
Kink Effect • Hysteresis and Latch • Parasitic Bipolar Transistor (PBT) • Transient and History Effects

4.8 Fully Depleted SOI MOSFETs **4**-29
Threshold Voltage • Subthreshold Slope • Transconductance • Meta-Stable Dip • Volume Inversion • Transition from Partial to Full Depletion

4.9 Scaling Trends **4**-33
Short Channels • Narrow Channels • Channel Thickness • Mobility Issues • Ultra-Thin Gate Dielectrics • Novel BOX

4.10 Multiple-Gate SOI MOSFETs **4**-40
Double-Gate MOSFETs • Triple-Gate MOSFETs • Gate-All-Around MOSFETs • Four-Gate FET

4.11 MEMS and Photonic Applications of SOI **4**-45

4.12 Conclusions **4**-45

Acknowledgments **4**-46

References **4**-46

Sorin Cristoloveanu
Institute of Microelectronics, Electromagnetism and Photonics

George K. Celler
Soitec USA

4.1 Introduction

Silicon-on-insulator (SOI) technology was initiated almost half a century ago for the fabrication of radiation-hard circuits. During 1970s and 1980s several SOI materials and structures were conceived for dielectrically separating the thin, active device volume from the silicon substrate [1,2]. The background idea is that in a bulk silicon MOS transistor, only a superficial layer approximately 100-nm thick is actually useful for electron transport, whereas the substrate causes undesirable effects. An example of a transistor made in SOI is shown in Figure 4.1 [3].

The overwhelming success of bulk-Si CMOS confined SOI technology to niche applications until late 1990s. Then several factors have increased the interest in SOI: invention of new fabrication methods for SOI materials (Unibond, ELTRAN, ITOX–SIMOX) and their optimization, need for lower power and higher speed circuits, emerging limitations of bulk CMOS scaling. Silicon-on-insulator transistors are now unchallenged in extending the frontiers of ultimate CMOS scaling.

This chapter is designed to offer a comprehensive tutorial on state-of-the-art SOI technologies. Section 4.2 introduces the basic concepts. Section 4.3 describes, in some detail, methods of fabricating conventional SOI substrates, while Section 4.4 includes novel modifications of SOI that further improve device performance. In Section 4.5 and Section 4.6, physical and electrical characterizations of SOI wafers are described. The physical mechanisms involved in the operation of fully depleted (FD) and partially depleted (PD) SOI MOSFETs are discussed in Section 4.7 and Section 4.8, respectively. Section 4.9 and Section 4.10 are dedicated to more advanced concepts related to the MOSFET miniaturization. We show that SOI is the *necessary* technology for CMOS scaling [3,4]. The SOI MOSFET is the smallest device conceivable in the CMOS world. We focus on the unusual dimensional effects that may take place in a small volume. The short-channel, narrow-channel and thin-body effects are addressed by including the impact of the gate and buried insulators. We finally show that 3D coupling effects govern the operation and optimization of advanced multiple-gate MOSFETs.

As soon as SOI wafers entered the "toolbox" of device engineers, many applications other than high performance fully scaled CMOS also emerged, such as BiCMOS for power and high voltage devices [5] and circuits for high temperature environments [6,7]. Also more options became available for circuits that are resistant to ionizing radiation [7,8].

Section 4.11 covers the use of SOI wafers in the area of micro-electro-mechanical systems (MEMS) and for the emerging microphotonic chips. Micro-electro-mechanical system applications of SOI take advantage of mechanical properties of the monocrystalline films, which are superior to those of polycrystalline Si [9]. Currently there is also much interest in utilizing SOI in various photonic applications. These usually rely on high refractive index contrast between Si and SiO_2, which permits photon confinement in small waveguides etched out of the top monocrystalline Si film [10].

For a reader that is interested in further details on SOI material fabrication and device physics, complementary useful information can be found in our previous review article [3].

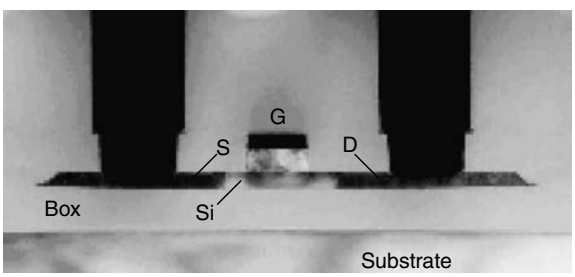


FIGURE 4.1 Cross-section of a thin-film SOI MOS transistor.

4.2 SOI Basics: A Tutorial

4.2.1 What It Is: A Definition

The most general definition of SOI involves a structure that consists of a layer of monocrystalline Si on an insulating dielectric film that is placed on a support "handle" substrate [1–3]. A predominant implementation of SOI, and some would argue the real definition is a monocrystalline Si film on amorphous SiO_2 on a single crystalline Si substrate. The fact that the top Si layer must be monocrystalline, but separated by an amorphous insulating film from the single crystalline handle substrate, poses a significant technical difficulty. Epitaxial deposition methods require a crystalline template; therefore it is not possible to grow a suitable device layer directly on SiO_2.

In terms of production volumes, 200 and 300 mm Si wafers predominate, with a layer of about 150 nm of thermally grown SiO_2 separating the single crystalline handle wafer from a single crystalline Si film that is 100 nm or less in thickness. Such wafers are primarily used for advanced CMOS applications. Somewhat thicker Si and SiO_2 films are needed for power and high voltage devices, for photonic applications and for some MEMS structures. These non-CMOS devices are often fabricated in older facilities that currently employ smaller wafer diameters (4, 5, and 6-in.).

In addition to the applications mentioned above, that rely on what is known as "thin SOI," where thin is anything with the Si film lesser than 3–5 μm in thickness, there are some niche applications, primarily in the MEMS field, that utilize Si layers from 5 to 100 μm thick.

Currently (year 2007), 80%–90% of the SOI substrates for the advanced electronic applications produced worldwide are made by Smart Cut™ technology that utilizes wafer bonding and layer transfer from a "donor" or "seed" wafer to a support handle wafer. The remaining 10%–20% of wafers are made by either in situ buried oxide (BOX) synthesis from implanted oxygen ions (SIMOX process) or by wafer bonding followed by mechanical grinding and etching (BESOI). Sometimes, in a process known as ELTRAN, an electrochemically formed porous Si layer is used to facilitate layer transfer after bonding of two wafers. In the main body of this chapter we focus our attention primarily on commercially relevant fabrication methods, but we also list numerous other approaches that were tried in the past.

4.2.2 How It Is Made: A Very Brief Description

Here we summarize the most common ways of making SOI wafers (see also Table 4.1). More details and the physical mechanisms are provided later in the chapter.

4.2.2.1 Smart Cut

Smart Cut™ technology relies on transfer of a high quality layer from one wafer to another. In typical applications, a thin layer of Si that is coated with thermal oxide is moved from its original donor or seed wafer to another Si wafer that is coated with only native oxide. The process sequence starts with oxidizing the donor wafer, followed by implanting ions, typically hydrogen, through the oxide and into silicon. Then the donor wafers and the handle wafers are very carefully cleaned to remove any particles and to provide surface chemistry that is most favorable to wafer bonding. Wafer pairs are properly positioned with respect to each other, lightly pressed together and the fusion wave spreads and makes the wafers stick together. In the following step, wafer pairs are split along the hydrogen-weakened zone that was previously produced within the donor wafers. Typically thermal energy provides a sufficient impetus for the split to occur, but mechanical force is used in some cases instead of heating. After the split, some finishing procedures are used to make the surface of the newly made SOI wafers smooth, and an annealing step strengthens the bonded interface. The donor substrates, which were split off the SOI wafers and are now thinner by a micron or less, are repolished in preparation for making another batch of SOI wafers.

4.2.2.2 Bond and Etchback

Bond and etchback SOI (BESOI) consists of bonding together of two oxidized and properly cleaned silicon surfaces, followed by a combination of mechanical grinding and chemical etching of a very significant fraction of one of the wafers in the two-wafer sandwich. By its very nature, this process is useful when the final Si film should be ≥ 5 μm, and is not practical for thin films.

4.2.2.3 ELTRAN

ELTRAN, which stands for epitaxial layer transfer, is another process that is based on wafer bonding. It utilizes a layer of porous Si formed by well-known electrochemical processes as the weak zone that facilitates the transfer of a Si film from the donor wafer to a new handle wafer. In order for this process to work, the top surface of the porous region has to be sealed by thermal annealing in hydrogen to provide a template for epitaxial Si growth.

After the epitaxy and oxide formation, wafer pairs are bonded, and then a fine water jet is used to cut through the weak porous layer. Surface finishing steps complete the process.

4.2.2.4 SIMOX

SIMOX is an acronym for *separation by implantation of oxygen*. This method relies on synthesis of BOX inside silicon by means of implanting a sufficient density of oxygen ions under a superficial layer of crystalline Si. The main challenge is placing enough oxygen under the surface of Si, while preserving the single crystalline nature of the superficial silicon. Temperatures of the order of 500°C–600°C during implantation prevent amorphization of the superficial layer of silicon. Much higher annealing temperature, greater than 1300°C, subsequent to implantation causes segregation of the implanted oxygen into a continuous buried layer of SiO_2.

4.2.3 How SOI Devices Work

An SOI circuit is composed of single-device silicon islands, dielectrically isolated from each other and from the underlying substrate. The *lateral* isolation offers compact design and simplified technology, whereas the *vertical* isolation via the buried insulator (BOX) allows achieving thin films (Figure 4.1), and eliminates most of the detrimental substrate effects (latch-up, punch-through, etc.).

Since the source and drain regions extend down to the BOX, the junction surface, leakage current, and junction capacitance are naturally reduced. The implications are improved speed, lower power dissipation, and higher temperature of operation [1,2].

In a typical MOS–SOI transistor (Figure 4.1), the current flows between source and drain. It is controlled primarily by the gate and secondarily by the substrate bias that acts as a *back* gate. Two inversion channels, front and back, can be formed, the main at the front Si–SiO_2 interface, and the other at the Si–BOX interface. They are independent in PD MOSFETs (Figure 4.2a) and coupled in FD MOSFETs (Figure 4.2b). The very thin drain and source regions make SOI devices to be less affected by short-channel effects, originated from "charge sharing" and from drain-induced barrier lowering (DIBL) [3,4]. Silicon-on-insulator MOSFETs are also rather tolerant to transient radiation effects.

Silicon-on-insulator circuits present a definite advantage in the highly competitive domain of high performance microprocessors. They also offer very beneficial characteristics for low-power and low-voltage circuits, where a small gate voltage shift is needed to switch the transistors from off- to on-state. Fully depleted MOSFETs benefit from a quasi-ideal subthreshold slope (60 mV/decade at room temperature). Operation at *equal voltage* consistently shows about 20%–30% increase in performance for SOI as compared to bulk silicon circuits. Conversely, operation at *equal low-power* dissipation may yield more than 100% performance gain. It is recognized that SOI circuits perform like bulk-Si circuits from the *next* technology generation. Even more importantly, ultra-thin film SOI MOSFETs have the unique capability of length scaling beyond the frontiers of bulk CMOS, thus enabling the Moore's law to extend further.

In the past, SOI circuits were handicapped by the cost of SOI wafers and by the unavailability of dedicated libraries for design. Silicon-on-insulator technology was then utilized only for circuits with

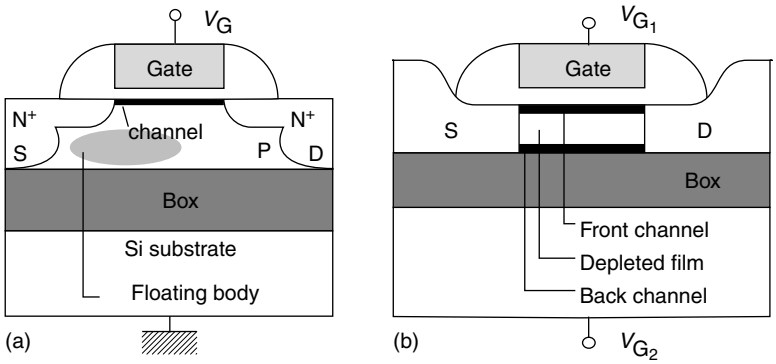


FIGURE 4.2 Basic architecture of (a) partially depleted (PD), and (b) fully depleted (FD) CMOS transistors on SOI.

high added value. But, in the late 1990s, the range of SOI applications has broadened considerably. Also, SOI has been recognized in the *International Technology Roadmap of Semiconductors* (ITRS) as the best option for many families of CMOS circuits.

Currently, high-end *microprocessors* are being fabricated on SOI (IBM, AMD, Freescale, Sony, etc.). RF SOI devices also show unchallenged capability. Extremely low-power circuits for mobile communication, portable processors operated with one-battery supply (0.5–1.2 V), and even battery-free systems (for watches) are currently fabricated with SOI technology. The SOI versatility has been taken advantage of, for conceiving capacitor-less DRAMs. Globally, SOI is an ideal substrate for systems-on-chip (SOC). The FD CMOS/SOI circuits can still operate successfully at temperatures beyond 300°C (for aeronautics and automobiles): the leakage currents are much smaller and the threshold voltage is less temperature-sensitive than in bulk Si [11]. Silicon-on-insulator circuits for the space industry can also be radiation-hard.

High voltage SOI devices take advantage of the dielectric isolation. Power transistors can have lateral configuration (for example, LD-MOSFETs for lighting applications) or vertical configuration. A vertical power device (IGBT, LDMOS, VMOS, etc.) can be integrated in a thick SOI film *or* in the non-SOI section of mixed SOI-bulk wafers. They can be rendered "smart" by being located next to a low-power SOI CMOS circuit [12].

Innovative device architectures and functions are based on the unique SOI features such as the adjustment of the thickness of the Si overlay and BOX, and the implementation of additional layers underneath the BOX. The 3D circuits containing consecutive thin silicon and BOX layers have been demonstrated with epitaxial lateral overgrowth (ELO) and zone melt recrystallization (ZMR) methods (See Table 4.1 for definitions of ELO and ZMR). For example, a 3D image-signal processor contains: arithmetic units and memories in bulk Si bottom level, fast A/D converters in an intermediate SOI layer, and photodiode arrays in an upper SOI layer [13]. The SOI family also includes optical devices: switches, waveguides, and modulators.

Silicon-on-insulator is an ideal material for *microsensors and MEMS*, where thin membranes are required; the interface between the BOX and substrate is a perfect etch-stop plane. Many transducers for detection of pressure, acceleration (airbags), gas flow, temperature, radiation, magnetic field, etc., have successfully been integrated on SOI [1,14].

Most of the Si *nanoelectronic* devices (SET, tunneling transistors, quantum dots and wires) have used SOI, either for process simplicity or for ultra-thin film capability [15,16]. Since quantum and tunneling effects become usual in standard SOI MOSFETs, the idea is to take advantage of them by importing device concepts from III–V semiconductors into the Si family. Resonant tunneling transistors can be accommodated in ultra-thin SOI films. In addition, band-to-band tunneling diodes (N^+–P^+) exhibit current–voltage characteristics with negative resistance (N-shape characteristics). Silicon-on-insulator

TABLE 4.1 Different Methods of SOI Formation

Method	Description
DI—dielectric isolation [78]	Oxide-isolated "tubs" of monocrystalline Si supported by a polycrystalline "handle" wafer.
SOS—Si-on-sapphire [79]	Si film epitaxially grown on sapphire substrates.
SOZ—Si-on-zirconia [80]	Si film epitaxially grown on ZrO_2 substrates.
Recrystallization from the melt:	Rapid melting of polysilicon films deposited over SiO_2 layer grown on a Si wafer, followed by controlled crystallization in a strong temperature gradient:
(a) Laser-seeded [81]	Cw laser beam raster-scanned across the surface, with via holes that connect the polysilicon film with the single crystalline substrate.
(b) Laser-unseeded [82]	As above, but no seeding vias in the oxide.
(c) ZMR—zone melt recrystallization with a hot wire [83]	A long and narrow molten zone is swept once across the entire wafer.
(d) LEGO—lateral epitaxial growth over oxide—stationary lamp heater [84]	A thick Si film is melted simultaneously across the entire wafer. Gradients due to seeding vias control crystallization.
ELO—epitaxial lateral overgrowth [85]	Selective Si epitaxial deposition, starting from via holes in SiO_2 and spreading laterally over the oxide.
SPE—solid phase epitaxy [86]	Oxidized Si wafers with via holes through the SiO_2 are coated with amorphous Si, which is epitaxially crystallized.
FIPOS—full isolation with porous oxidized silicon [87]	Porous Si is formed locally under islands of crystalline Si, then it is oxidized to form isolation.
Heteroepitaxy of crystalline insulators, followed by single crystalline Si [88]	CaF, ZrO_2, spinel, Al_2O_3 and other crystalline insulators have been used.
SIMOX—Separation by IMplantation of OXygen [61]	BOX layer is synthesized in situ from implanted oxygen.
Wafer bonding and etch-back [25]	Two wafers are bonded with an oxide layer in between. One of the wafers is thinned by grinding and etching.
Smart-Cut™ process-layer transfer facilitated by ion implantation [39]	One wafer is implanted, typically with H or noble gas ions. The Si layer above the implanted region is transferred to a handle wafer by wafer bonding and splitting along the implanted region.
ELTRAN process-layer transfer facilitated by porous silicon [58]	Epitaxial layer is grown on a porous Si region and transferred by bonding and splitting to a handle wafer.
SON—silicon-on-nothing [89]	Successive epitaxy of SiGe and Si films on a Si substrate is followed by removal of the sacrificial SiGe, which leaves lithography-defined small cavities. The cavity walls can be coated with SiO_2.

has the merits of reducing the parasitic leakage current and using the back-gate biasing, which modifies the local carrier concentration.

The fabrication of tiny Si islands for *Single-Electron Transistors* is easier in very thin SOI. For example, a SET inverter is composed of nanometer-scale islands formed by anisotropic sacrificial oxidation combined with local stress effects (lower oxidation rate on the sidewalls). Two such SET islands, properly interconnected and biased, give rise to an inverter with gain higher than unity [16].

4.3 SOI Wafer Fabrication Methods—Some Details

Two fundamentally different approaches to SOI wafer fabrication are in commercial use. The predominant one is based on bonding of two wafers together. Afterwards one of the substrates in the bonded wafer pair must be transformed into a thin film of uniform thickness, low stress, and excellent crystallinity. The thinner the required film, the more difficult is this task. Another basic approach to SOI fabrication, known as SIMOX, involves direct synthesis of SiO_2 inside a Si wafer.

First we describe the science and technology of bonding of two wafers together, as this is generic to all wafer-bonding-based approaches. The different paths from a bonded wafer pair to a finished SOI wafer are explained next. Then the details of the SIMOX method, which does not include bonding, are given. And for completeness and historical background, at the end of this section we list a variety of other methods that are scientifically interesting, but that have not led to any significant commercial activity, and also two methods that predate what we now define as SOI and that have lost much of their commercial importance.

4.3.1 Wafer Bonding

Just like two glass plates can stick together, properly cleaned Si wafers can be bonded simply by pressing them together at one point. Clean and hydrophilic-oxidized Si surfaces stick together because of attractive forces between water molecules trapped at the interfaces. This initial bonding is relatively weak, of the order of 0.1 J/m^2. A subsequent heat treatment causes outdiffusion of water molecules from the bonding interface region, their dissociation followed by reaction with Si at the Si/SiO_2 interfaces that leads to a slight increase in the SiO_2 thickness. Several excellent reviews on the science and technology of wafer bonding have been published [17–20]. A chapter by Haisma in Ref. [20] details the rich history of patents related to bonding of wafers and of other structures [21]. It includes other types of bonding, such as anodic bonding that has many important applications, but is unsuitable for making silicon-based electronic devices. Here we only consider direct wafer bonding that is also known as fusion bonding.

The nature of bonding forces depends on the materials being placed in contact and on the surface preparation (surface chemistry). Van der Waals forces will hold together two very flat and very clean surfaces, for example, in ultra-high vacuum environment. These forces depend on polarizability of atoms or molecules on two surfaces placed very close (<1 nm) to each other. Spierings et al. [22], have estimated the surface bond energy of 0.0075 J/m^2 for two fused silica surfaces held together by Van der Waals interactions. Direct bonding of two clean silicon surfaces, which is gaining some technological importance, would likely rely on Van der Waals forces.

Formation of SOI wafers by wafer bonding relies on another mechanism, that of chemical assistance through hydrogen bonds and water molecules. After wafer cleaning in an RCA solution [23] followed by a water rinse, the surfaces are coated with OH groups, which attract water molecules—i.e., the surfaces are hydrophilic. Hydrogen bridges and water molecules initially hold two such surfaces together. The strength of "water" bonding of two oxidized Si surfaces is 0.10 J/m^2 according to Stengl [24]. The actual bond energy depends on the details of surface preparation, bonding conditions, and ambient.

The field of wafer bonding for SOI applications was initiated by Lasky at IBM [25]. He showed that bonding only required proper cleaning of the wafer surfaces followed by applying some slight mechanical force. Independently of Lasky and about the same time, Frye et al. [26] at Bell Labs showed that wafers could be bonded by pressing them together by means of electrostatic forces. Since Lasky's method did not require electric field, it became the dominant approach.

The evolution of IR transmission spectra for light that was propagated along the interface between two oxidized Si wafers was investigated in some detail by Weldon et al. as a function of thermal annealing [27]. They looked mainly at the case where both Si surfaces were coated just with a chemical (native) oxide that formed during a modified RCA cleaning step. In these experiments, spectra associated with Si–O, Si–H, and O–H stretching vibrations were detected and studied.

Based on the interpretation of the spectra, the following model of the bonding mechanism emerged. Immediately after the wafers are fused together at room temperature, there are a few monolayers of water trapped between two native oxide films, each approximately 4 Å thick. Heating samples to 300°C eliminates about 75% of water by diffusion through the thin oxides to the Si interfaces, where an additional 4 Å of oxide is formed. The oxidation reaction liberates molecular hydrogen. When the bonded wafer pair is heated further, up to 800°C, the remaining water diffuses away from the bond interface, so that interfacial hydroxyl groups on opposing surfaces can couple into bridging siloxane

bonds that begin to fuse the two wafers together:

$$\text{Si–OH} + \text{HO–Si} \rightarrow \text{Si–O–Si} + \text{H}_2\text{O}$$

The water produced in the reaction above causes further oxidation of Si and the liberated H is trapped in the newly formed oxide. After further heating of the samples up to 1100°C, a complete closure of the interface occurs by coupling of the remaining interface hydroxyl species and diffusion of the hydrogen into the Si bulk:

$$\text{Si–OH} + \text{HO–Si} \rightarrow [\text{Si–O–Si}]_{\text{strained}} + 2\text{H};\ \ \text{H} \rightarrow \text{H(silicon)}_{\text{bulk}}$$

The results for Si wafers with a thermal oxide instead of a chemical oxide are similar, but about 40% less water is trapped initially between wafers—possibly because the surface of the thermal oxide is smoother.

In order to systematically measure the strength of the bonded interface between two wafers, Maszara et al. [28] introduced a double cantilever or crack opening method. In this approach, a wedge is pushed from one side between two bonded wafers and the length of the debonded area is correlated with the wedge thickness and elastic coefficients of the wafers that are being separated. Measurements showed that the bond energy immediately after bonding can vary across a large range for two oxidized Si wafers [22]. However, after 50–100 h in room ambient, the bond energy usually saturates at a value of 130 ± 4 mJ/m^2. This saturation is caused by hydrogen bridges that gradually fill gaps between two somewhat imperfect surfaces [29]. After the initial room temperature bonding, annealing at elevated temperatures fuses the wafers. The bond strength reaches a maximum value of approximately 2 J/m^2 after a heat treatment at 1100°C [28].

In a recent paper, a quantitative description of the dynamics of the adhesion process was proposed and compared with experiments [30]. The model is consistent with data and with literature results. On this basis, the authors suggest that the bonding front velocity can become a method for nondestructive determination of the bonding energy.

Exposing either one or both surfaces to oxygen plasma before bonding them together can significantly enhance the room temperature bonding energy [31–33]. Approximately the same enhancement can also be achieved with nitrogen or argon plasma. The mechanism of the enhancement appears to be related to the plasma-induced surface damage and not to surface-trapped charges, as some initially postulated. The damage increases surface porosity, which accelerates outdifusion of water from the interface between two bonded wafers.

4.3.2 Smart Cut™ Technology

Michel Bruel of LETI discovered that the blistering effect, well known [34–37] in the ion implantation community and also in the nuclear reactor field, could be harnessed to produce very well-controlled transfer of thin layers from one semiconductor substrate to another. A patent on forming SOI films by controlled exfoliation was filed in 1991 [38], and technical publications on this subject started appearing in 1995 [39,40]. This important discovery led to a new way of making SOI wafers and greatly accelerated transition of SOI into the commercial arena.

Bruel's method, utilizes various gas ions—most commonly, hydrogen—for implantation as an atomic scalpel that cuts through Si wafers. Hydrogen ions, when implanted to a dose of greater than 5×10^{16} cm^{-2}, produce fine microcavities in the Si lattice. Some hydrogen ions attach themselves to the dangling Si bonds in the microcavities, while others fill these voids. If such an ion-implanted wafer is heated up, typically to 400°C–500°C, more hydrogen segregates into the voids in the form of molecular hydrogen, H_2, the pressure builds up to a point of fracture, and the surface of Si becomes pockmarked with blisters. This is clearly an undesirable effect of ion implantation. For an implant dose exceeding approximately 10^{17} cm^{-2}, blistering may occur even without the heat treatment. Blistering phenomena caused by surface bombardment with hydrogen or inert gases have been seen in the past. Extended defects

produced by lower doses of hydrogen in silicon were detected and analyzed by Johnson et al. [41] and by Cerofolini et al. [42]. Johnson et al. even reported observing hydrogen platelets and microcracks [41].

The essence of Bruel's invention was the realization that the effect, which was always considered deleterious, could be utilized to produce a highly fragile planar zone that would facilitate controlled cutting through the crystalline lattice. The most important ingredient of Bruel's method was a stiffener placed on the top of the implanted wafer. This thick and stiff layer prevents blistering and redirects the pressure that builds up in microcavities in a lateral direction, as shown schematically in Figure 4.3. After the stiffener is attached to the suitably implanted wafer, some form of force is applied to split the crystal. Heating of the wafer builds up pressure within the weakened plane that can cause a split. Alternatively, application of a mechanical or other stress will also cause cleavage of the crystal.

4.3.2.1 Process Description

The idea of controlled exfoliation and layer transfer was developed into a commercial process for SOI that is known as the Smart Cut™ technology. In this method, a typical sequence of steps required to make SOI wafers is shown in Figure 4.4. A seed wafer, from which a layer of Si will be removed, is oxidized to a desired thickness. This oxide will become the BOX after bonding. The next step is hydrogen implantation through the oxide and into Si with a dose that is typically greater than $5\times10^{16}\,\mathrm{cm}^{-2}$. After the implantation, the seed wafer and the handle wafer are carefully cleaned in order to eliminate any particles and surface contaminants, and to make both surfaces hydrophilic. Wafer pairs are aligned and contacted so that the fusion wave can propagate across the entire interface [30]. A batch of bonded wafer pairs is loaded into a furnace and heated to a temperature of 400°C–600°C, at which point the wafers split along the hydrogen implanted plane. The as-split wafer surface has a mean roughness of a few nanometers. A light touch-polish or other surface treatment brings the same surface roughness as in the standard bulk Si

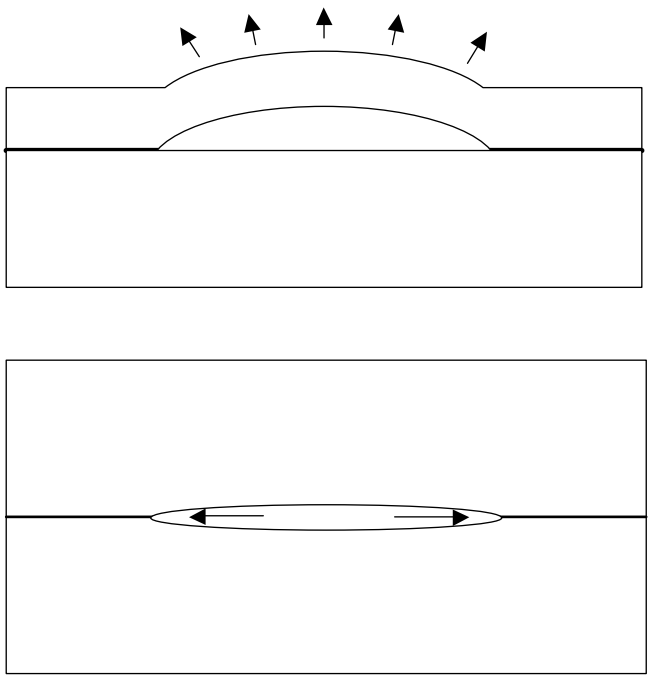


FIGURE 4.3 A schematic drawing illustrating the importance of a "stiffener" layer. At the top a microcavity filled with gas, typically hydrogen, forms a blister that is about to burst. At the bottom, the stiffener redirects the pressure in the lateral direction.

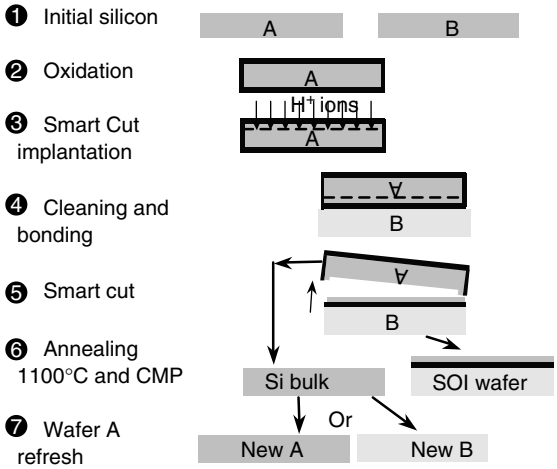


FIGURE 4.4 Smart Cut schematic.

wafers, i.e., $R_a < 1$ Å across 1×1 μm^2. The seed wafer is reclaimed and, if necessary, repolished so that it can be used again.

There are several important practical aspects to the method of controlled transfer of a layer of Si, the thickness of which is approximately defined by ion implantation energy, to a handle wafer. This approach makes it possible to reuse the seed wafer several times, thus reducing the final cost of the SOI wafer. It is the premium seed wafer that defines the quality of the SOI film, whereas the handle wafer only serves as a mechanical support and can have less stringent parameters. Defining film thickness by implantation energy leads to a much better thickness uniformity control than is possible with either mechanical or chemical thinning. For that reason BESOI techniques are typically limited to films thicker than 5 μm, where the absolute thickness control is easier. The thickness of the silicon film and/or BOX can be adjusted in a wide range in the Smart Cut™ process by tuning the implant energy and oxidation time. The thickness of the silicon film in current applications typically runs from about 5 nm to 1.5 μm. The thickness of the silicon dioxide is typically set at 5 nm to 5 μm. Unibond® wafers are thus adaptable to most device architectures, from ultra-thin CMOS to thick-film power transistors and sensors. It is also worth noting that only conventional equipment is needed for mass production of 8 and 12″ wafers.

4.3.2.2 Hydrogen-Induced Splitting

In this section, we describe the physical mechanisms that come into play when ion implantation into silicon generates certain types of damage that eventually lead to splitting of a thin layer away from the main substrate. This splitting can be achieved by implanting hydrogen or noble gas ions. Here, we will limit the discussion to hydrogen, as its interactions with silicon are best understood and most experimental data are available for this case [43]. Later we will also consider the co-implantation effects of hydrogen and helium [44,45].

High-dose hydrogen implantation into Si produces platelets that tend to trap hydrogen. For H^+ implant doses greater than 2×10^{16} cm^{-2} at 60 keV, the initial platelet density is high enough so that some of them grow through the thermal cycle at the expense of smaller ones that are dissolved. This is a typical case of coarsening of the structure by means of Ostwald ripening. For H^+ doses lesser than 2×10^{16} cm^{-2} at the same implant energy, the platelet concentration in a standard silicon wafer is usually insufficient to trap H for extended times at elevated temperatures. Low platelet density and their small size are not favorable for producing microcracks.

4.3.2.2.1 Implanted Silicon Wafers—No Annealing

Aspar and his associates analyzed in detail the as-implanted Si wafers by transmission electron microscopy (TEM) [46,47]. Additional information on hydrogen chemical bonds, inside the silicon matrix, was obtained by infrared absorption spectroscopy in the multiple internal transmission configuration [48]. After implantation of hydrogen doses in the range of a few 10^{16}–10^{17} H^+ cm^{-2}, there are no defects at the Si surface, but XTEM micrographs indicate the presence of platelets or microcavities confined around the maximum hydrogen concentration depth, R_p. These platelets are visible in high resolution TEM and when using bright field conditions far from any Bragg diffraction. For an approximately 6×10^{16} H^+ cm^{-2} implantation dose, the platelets are about 1–2 atomic planes in height and about 10 nm in diameter [49]. In (100) wafers, most of the platelets lie on the (100) planes parallel to the surface, and to a lesser extent on the (111) planes. Recently some platelets that are perpendicular to the (100) surface were also observed [50]. The platelets can be observed by TEM before any heat treatment, and their diameter grows during annealing, while the density goes down and the total volume stays approximately constant.

By coupling infrared light into as-implanted wafers [51], Aspar et al. obtained spectra, which showed that hydrogen is bound to silicon mainly as monohydride species [48]. Si–H bonds at (111) surfaces and Si–H_2 bonds at (100) surfaces can also be observed. They correspond to the presence of (100) and (111) platelet-like defects observed by TEM in the implanted region. The size of the (100) microcavities depends on the implantation conditions and can vary from small platelets up to microcracks. In some extreme cases, these microcracks located around the projected range R_p can have a size of several microns, and even lead to formation of the blisters that are optically visible at the wafer surface. For instance, a high dose hydrogen implantation of 2×10^{17} H^+ cm^{-2} leads through growth of the platelets to the microsplitting phenomenon.

4.3.2.2.2 Annealed Hydrogen-Implanted Silicon Wafers

In wafers implanted with a few 10^{16}–10^{17} cm^{-2} hydrogen ions, heating induces growth of microcavities and results in the formation of microcracks, as shown in Figure 4.5. A quantitative TEM study of the thermal evolution of the platelets requires the use of specific imaging conditions [46,47], so that the size distribution, density, and overall volume occupied by the cavities can be measured as a function of the

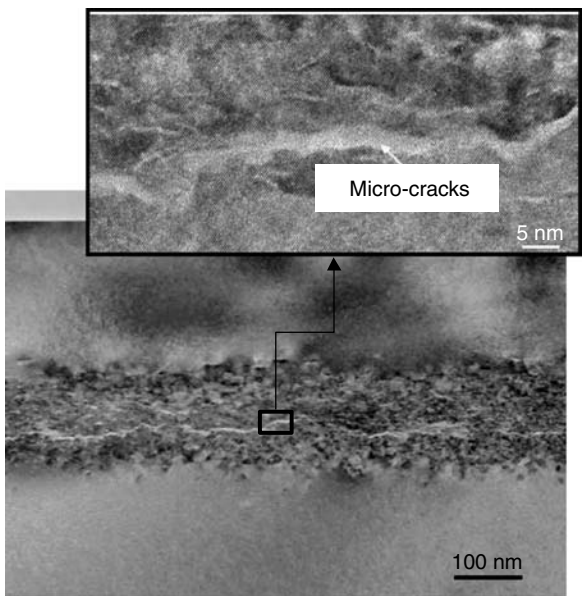


FIGURE 4.5 Microcracks in a "weakened" zone after hydrogen implantation and thermal treatment.

thermal history. For 6×10^{16} H^+ cm^{-2} samples implanted at 70 keV, the first minutes at 450°C cause the growth of the (100) microcavities that are parallel to the surface, while their density decreases and the total volume remains approximately constant. This is consistent with an Ostwald ripening mechanism for growth of larger microcavities at the expense of smaller ones that shrink and disappear. In the process, hydrogen diffuses from the small cavities to the larger ones. The results are in agreement with an earlier study of a wafer implanted with a lower, 3×10^{16} cm^{-2}, hydrogen dose [47].

Splitting kinetics for different crystalline orientations were first reported by Zeng et al. [52], and investigated in more detail by Bourdelle et al. [53]. Arhenius plots obtained by the latter for splitting along (100), (110), and (111) planes are presented in Figure 4.6.

4.3.2.3 Hydrogen and Helium Co-Implantation

A weakened zone can be produced inside a Si crystal not only with hydrogen, but also with other implanted species, in particular with helium and other noble gases. However, hydrogen, either alone, or in tandem with another species, is preferred because of its reactivity with the internal surfaces of a semiconductor. For example, a study of co-implantation of H^+ and He^+ has demonstrated that each plays a somewhat different role. Hydrogen interacts chemically with the implantation damage to produce platelet-shaped microvoids. He^+, implanted after the hydrogen, fills the cavities and provides most of the pressure that causes splitting of a Si film from the bulk substrate. Agarwal et al. [44], have demonstrated that splitting could be achieved with doses as low as 7.5×10^{15} cm^{-2} H^+ and 10^{16} cm^{-2} He^+. In contrast, under the same implantation and annealing conditions, He^+ alone requires a significantly higher dose of 2×10^{17} cm^{-2} and H^+ alone requires 6×10^{16} cm^{-2}. According to Agarwal, reversing the sequence of implants in the low-implant dose case does not provide the same benefits as the "hydrogen first" case. More recent evidence indicates that implantation conditions, and in particular the relative implantation depth for the two species, plays an important role. Lagahe-Blanchard et al. have shown that the advantage of H^+ implant first disappears, if the projected range R_p of He^+ ions is greater than that of H^+ ions [54]. They also demonstrated that co-implantation can improve the quality of transferred Si layers and provide better splitting kinetics. Increasing the total co-implant dose to 5.5×10^{16} cm^{-2} made

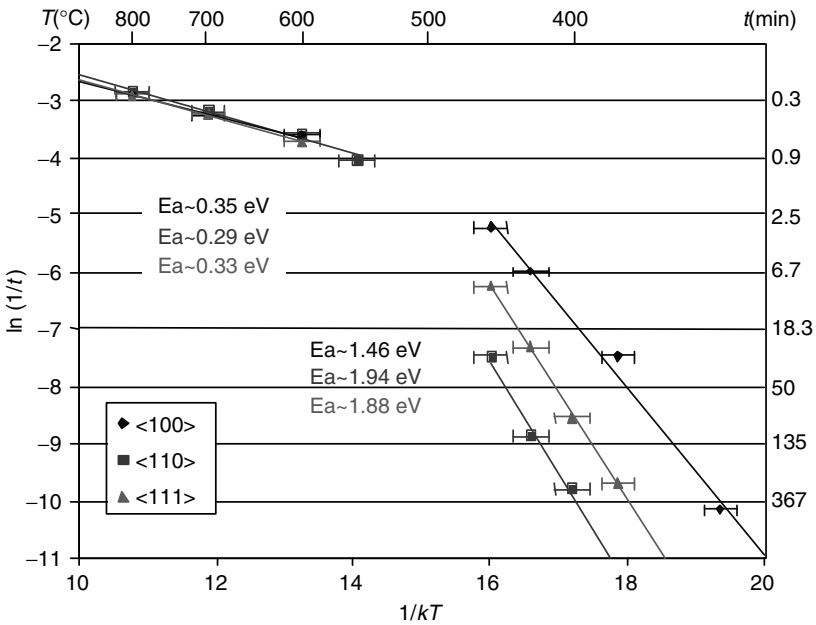


FIGURE 4.6 Arhenius plots for splitting along three crystal orientations: (100), (111), and (110).

possible thermal splitting at a significantly reduced thermal budget, a feature that has significant practical advantages, especially when heterostructures are made [54,55].

Nguyen et al. [45] investigated in great detail the damage formation caused by co-implantation as a function of the implant sequence and ion energy. Their analysis of TEM and IR spectra explains how subtle changes in amorphization trends for four sets of initial parameters lead to slightly different conditions for splitting and it shows how they affect the quality of as-split surfaces.

4.3.3 Bond and Etchback

Bond and etchback SOI, also known as BESOI, is very useful for forming thick Si films that are dielectrically isolated from the substrate [17,21,25]. Films thicker than 10 μm are relatively simple to make by standard grinding, etching, and polishing methods (see Figure 4.7). Without any specific thickness markers or etch-stop layers it is difficult to control film thickness and uniformity to less than 1 μm, but for thick film applications this degree of precision is often adequate.

Better thickness control can be obtained by introducing etch stop layers or at least depth markers. For example, if there is a heavily boron-doped buried layer many micrometers under the surface, it will sharply reduce silicon removal in a KOH etch bath. Thin SiGe layers or oxygen implants can play a similar role. One significant disadvantage of the BESOI process is that two Si substrates are always consumed to make one SOI wafer. Etch stop layers, if needed, require additional process steps, typically growth of thick epitaxial layers.

Bond and etchback SOI wafers are used in many MEMS applications, where 10–100 μm thick films are often required. Some high power and/or high voltage devices also take advantage of these structures.

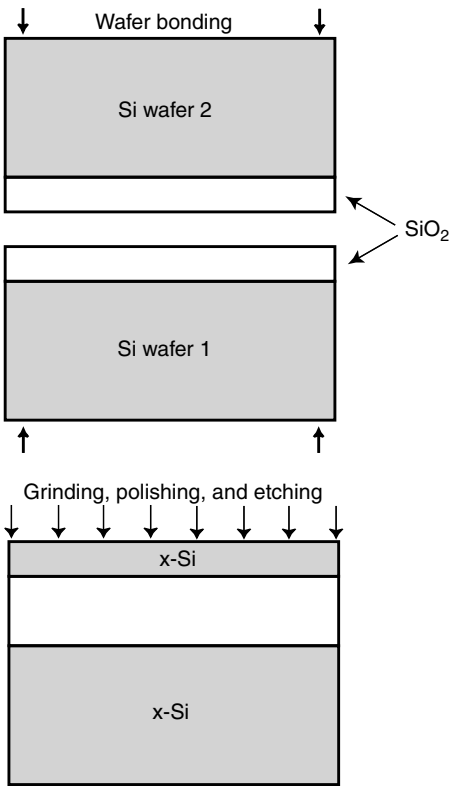


FIGURE 4.7 Bond and etch SOI (BESOI) schematic.

4.3.4 ELTRAN

In Section 4.2.2.3, we have already indicated that porous Si is the key to the ELTRAN process. Porous silicon is formed when Si is anodized in an HF solution, an electrochemical process that was first described in detail by Uhlir in 1956 [56]. In the electrolytic cell, a random network of nano-pores is etched out in Si. Baumgart et al. were the first to show that porous Si surface is still an acceptable template for epitaxial Si deposition [57]. Yonehara et al. demonstrated that it is possible and practical to grow device quality single crystalline Si films on the top of porous Si layers [58]. They utilized high temperature annealing in hydrogen, which greatly increases the mobility of Si atoms, to seal the pores at the surface, thus improving the quality of the substrate for the subsequent Si growth. After Si epitaxy and oxidation to form what will become the buried oxide, the donor wafer is bonded to a Si handle wafer. To complete layer transfer, the wafer stack is fractured mechanically along the porous Si layer that is mechanically weak. A fine water jet, aligned with the wafer plane and incident at the perimeter of the wafer pair that is rotated in front of it, breaks the wafers apart.

The ELTRAN team further improved the quality of as-split surfaces by modifying the physical properties of the porous layer [59]. The size and density of pores are a function of the current conditions during anodization. For best results, very fine pores are formed at the donor wafer surface, with a second coarser layer positioned deeper into the substrate, as shown schematically in Figure 4.8. The built-in interfacial stress at the boundary between these two porous layers facilitates water jet-induced cracking along the planar interface, leading to a more uniform cleavage.

The finishing steps for the newly formed SOI wafers consist of etching away of the residual porous Si followed by surface smoothing by a second application of hydrogen annealing at about 1100°C. The original donor wafer, from which a thin layer of porous Si was split-off, can be reclaimed, polished, and used again.

4.3.5 SIMOX

4.3.5.1 Basic Concepts and Early Results

In mid-1970s, Izumi and his coworkers at NTT started investigating SOI formation by direct synthesis of SiO_2 from implanted oxygen ions. Experiments on BOX formation were initiated with an Extrion 200-20a ion implanter [60]. In 1978, Izumi et al. demonstrated a 19-stage CMOS ring oscillator made in the new material, which they called SIMOX [61].

In these early attempts to form SOI, enough oxygen needed to be implanted to reach the stoichiometric concentration of oxygen in the buried SiO_2 layer already during implantation. To achieve a Si film of about 200 nm, 200 keV ion energy was necessary, and approximately 2×10^{18}

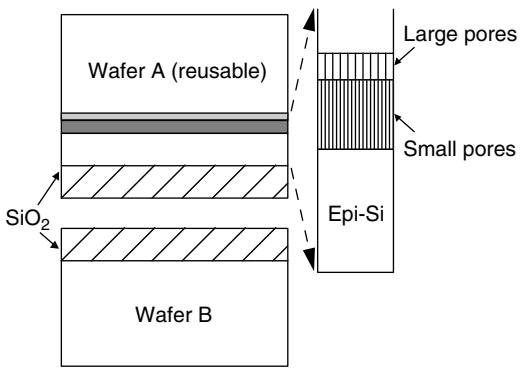


FIGURE 4.8 ELTRAN schematic.

oxygen ions cm^{-2}, a dose 100 times above a typical implant dose utilized in device processing. Because of this high dose, crystalline lattice damage was very extensive. In fact, at room temperature, the entire layer penetrated by the ions would have been completely amorphized. To preserve single crystalline nature of the Si overlayer, thermal annealing was needed concurrent with the implantation. At temperatures greater than 500°C, the dynamic annealing of damage during implantation leaves monocrystalline Si near the Si surface, where the ion energy is highest and thus there is less displacement damage. After implantation, very high temperature annealing is required to react oxygen ions with Si in order to form SiO_2 while annealing the damage in the Si layer above and in the Si substrate below the oxide. Evolution of the SIMOX structure during oxygen implantation is shown in Figure 4.9 [62].

4.3.5.2 High Temperature Annealing

Conventional furnaces with fused silica liners are limited to temperatures lesser than 1250°C. The microstructure quality and device yields after 1250°C anneals were poor. Annealing at very high temperatures was developed to improve the microstructure. Dissolution of oxide precipitates at elevated temperatures depends on Ostwald ripening that causes growth of the precipitates with a radius above a critical value at the expense of the smaller ones that are dissolved [63]. At temperatures T_a above 1300°C the critical radius approaches infinity—only the planar BOX remains. Annealing of SIMOX structures at 1300°C for several hours [63] or at 1405°C in a lamp furnace for 30 min [64,65] demonstrated the feasibility of forming atomically sharp and planar interfaces between Si and the buried oxide. The dependence of SIMOX microstructure on annealing temperature is well documented [66,67]. All SIMOX wafers that are currently produced undergo annealing at 1300°C or above (typically ~1350°C) in furnaces with either polysilicon or SiC tubes.

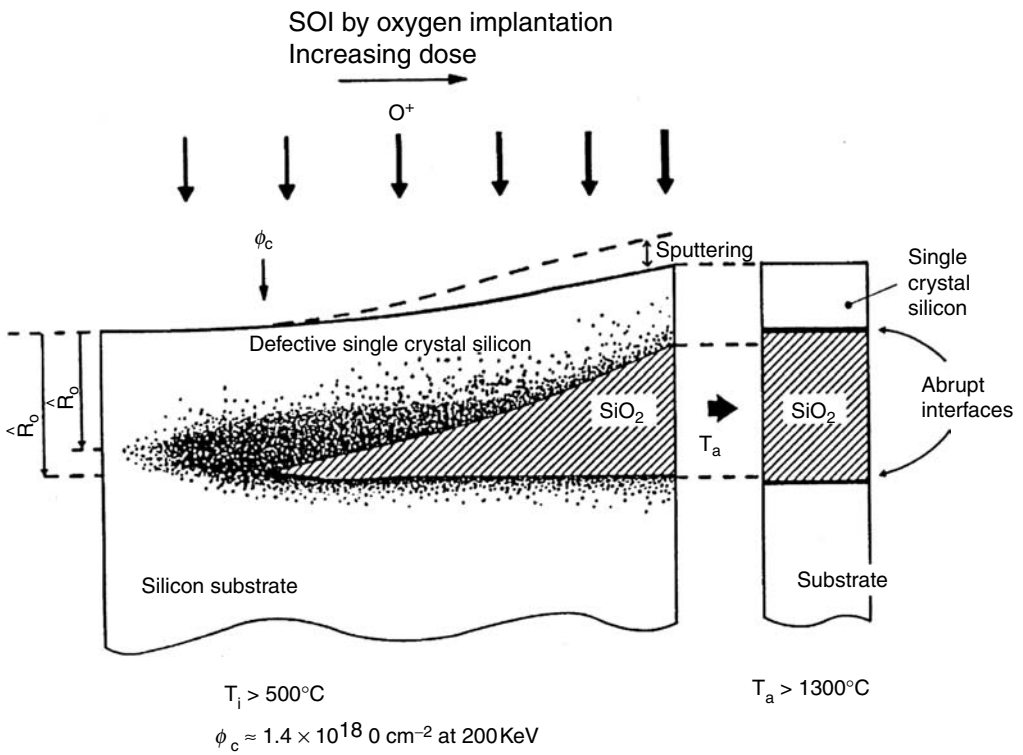


FIGURE 4.9 Evolution of a separation by implantation of oxygen (SIMOX) structure during oxygen implantation. (Adapted from Hemment, P. L. F., Reeson, K. J., Kilner, J.A., Chater, R. J., Marsh, C., Booker, G. R., Davis, J. R., and Celler, G. K., *Nucl. Instr. Meth. Phys. Res.*, B21, 129, 1987.)

4.3.5.3 Implant Optimization

Temperature T_i of the Si substrates during implantation also plays a critical role in achieving good microstructure of SOI wafers. In early SIMOX wafers, 10^{10} cm^{-2} threading dislocations were typical. Reduction to 10^6 cm^{-2} threading dislocations was achieved by increasing implantation temperature to approximately 600°C [68]. To further reduce the defects, the dedicated oxygen implanters had to be improved and optimized for very high doses. To avoid sputtering of metals from the walls of the implantation chamber and onto the wafers, the chamber interior was coated with silicon.

4.3.5.4 Low-Dose SIMOX

SIMOX technology has greatly evolved and improved since the first device demonstrations by Izumi et al. [61]. The efforts to reduce defect density and cut down the cost of processing were helped by the device-scaling trend that required moving to thinner films. Thinner Si film means lower implant energy, and thinner BOX translates into a lower implant dose. Lower oxygen dose, in particular, helps in improving crystalline quality of the wafers. Cutting the oxygen dose is not entirely trivial; mechanisms that allow formation of a planar and uniform BOX are not understood well enough to make theoretical predictions of optimum implantation conditions. Experimentally a few "sweet spots" were identified.

For example, a high quality planar BOX was obtained at a much lower dose of 4×10^{17} cm^{-2} by modifying the implant and anneal conditions [69]. This low-dose BOX is about 100-nm thick, suitable for the sub-0.25-μm CMOS devices. The feasibility of even thinner BOX films has been demonstrated with O^+ implantation of just 2×10^{17} cm^{-2} at 65 keV, followed by 4 h anneal at 1350°C, which resulted in a 56-nm thick oxide [70].

4.3.5.5 ITOX Process

The success of SIMOX with a thin BOX was greatly facilitated by a discovery and development of a procedure known as internal oxidation (ITOX). Synthesized BOX is prone to have some pinholes or Si pipes that electrically short the Si film to the substrate. However, when an SOI wafer is oxidized at approximately 1350°C, a small fraction of oxygen that diffuses through the surface oxide also diffuses through the silicon film (instead of oxidizing the Si surface) and reacts with it at the Si/BOX interface [71]. This internal oxidation improves the stoichiometry of the BOX, giving it properties much closer to those of thermally grown SiO_2, reduces or closes Si pipes, and as shown schematically in Figure 4.10, it slightly increases the overall thickness of the BOX.

Commercially available SIMOX wafers, known by trade names like MLD (modified low dose) and Advantox™, fine-tune the processing parameters and add some additional improvements. For example, the SIMOX structure was modified by adding a low dose (10^{15} cm^{-2}) room temperature implant after

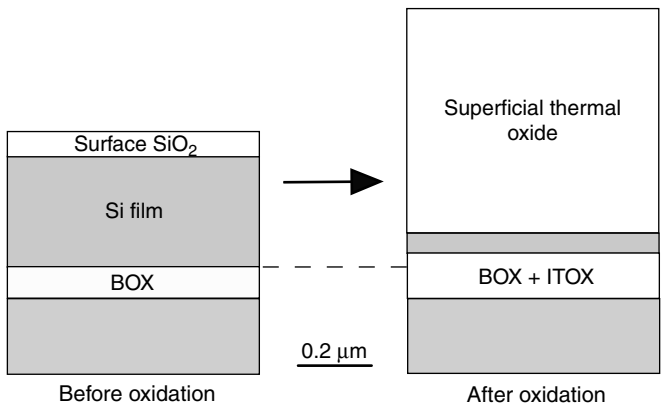


FIGURE 4.10 Schematic representation of the internal oxidation (ITOX) process.

the "standard" hot implant [72]. This additional step leads to a more planar BOX layer with fewer Si inclusions, since it amorphizes the Si just above the peak of oxygen concentration R_p.

4.3.5.6 Patterned Buried Oxide

SIMOX SOI substrates are normally formed by a blanket oxygen implantation. But there are some applications, such as SOC and microprocessors with embedded DRAMs, where it might be useful to have BOX only in some parts of the wafer, with other areas remaining as conventional bulk Si. Attempts to implant oxygen locally through a thick-patterned masking layer were started already in 1980s [73,74]. In these experiments, after the standard 2×10^{18} cm^{-2} oxygen dose was implanted, the boundary between the SOI and bulk regions was extremely defective because of high stresses at the oxide edges. After low dose SIMOX was developed, the density of defects at the boundary was reduced considerably, but it was still high enough to potentially cause problems with device yield and reliability.

Cohen and Sadana, reduced further the number of defects in the transition region between SOI and bulk [75]. They used a blanket low-dose oxygen implantation, followed by a touch-up amorphizing O^+ implant at room temperature and then by patterned internal oxidation (ITOX process). This led to a thicker BOX in regions exposed to oxidation (but also much thinner Si in the same areas). In the extreme case of a subthreshold O^+ implant dose of 1.5×10^{17} cm^{-2}, the BOX was discontinuous under the oxidation mask. Ogura [76] used a different approach in which oxygen precipitated into BOX layers only in regions that were previously damaged by He^+ implantation.

Researchers from Shanghai have recently reported interesting results on forming low-dose patterned BOX by carefully optimizing the dose and energy conditions, followed by 1300°C anneal in Ar with 3% O_2 [77]. Good results were obtained at 3.5×10^{17} cm^{-2} @100 keV and at 2×10^{17} cm^{-2} @ 50 keV.

4.3.6 Other Fabrication Methods

4.3.6.1 Overview—SOI Menagerie

Over the years, there have been many attempts to utilize localized templates and then extend epitaxial growth from these templates to other regions, but these approaches, although scientifically interesting, have not led to many practical solutions. The situation is somewhat different when the insulator is not amorphous, but monocrystalline. Heteroepitaxial growth of Si on a bulk crystalline sapphire, became a commercial technology known as SOS, but was found to be of limited utility except for rad-hard CMOS. Several other approaches have been studied, which significantly enlarged our body of knowledge about the microstructure and morphology of thin silicon films. These methods, including those already discussed by us, are summarized in Table 4.1 that is reprinted from Celler and Cristoloveanu [3], and an interested reader can obtain the details by following the references listed in the Table.

The great variety of approaches did not lead to commercial applications. Many of these techniques require intimate integration of SOI fabrication with device making—in other words, the SOI is patterned and the pattern is device- or circuit-dependent. This means that the SOI fabrication needs to be completed by the same production line that makes devices. Complexity of doing both discouraged most semiconductor chipmakers.

There were two interesting exceptions in the past, DI and SOS. And to complete the picture, there is also an interesting recent approach, silicon-on-nothing (SON), in which the isolation layer is formed as a part of the device fabrication process. All three methods are briefly explained below.

4.3.6.2 DI Technology

This technology was developed at TI by Bean and Runyan for bipolar and high voltage applications [78]. The approach required one starting bulk Si wafer, patterning and deep etching, thick oxide, and deposition of very thick polysilicon at high temperatures.

The technology provided complete dielectric isolation between large regions of single crystalline Si. It was suitable for some large devices and high voltages. The detrimental aspect was the presence of a

polycrystalline handle wafer—dimensional stability (and with it wafer flatness, bow, and warp) was inferior to conventional Si wafers.

LEGO process was an attempt to preserve the unique structure of DI wafers while providing a single crystalline Si handle wafer [84]. A related development to produce power devices in very thick films is ongoing [90].

4.3.6.3 SOS

Heteroepitaxy played an important role in the early days of SOI. SOS [79] is about as old as DI, and the main motivation for its development was radiation hardness. Since the crystalline quality of as-grown epitaxial Si on Al_2O_3 is rather poor, post-epi steps are usually added. Commonly they include a Si implant that amorphizes the most defective Si near the alumina interface. In the subsequent solid phase epitaxial regrowth, the template is the near-surface Si that has better crystalline quality.

4.3.6.4 SON

Silicon-on-nothing process consists in growing by selective epitaxy a sacrificial SiGe layer in STI-predefined regions of a bulk-Si wafer. A silicon film with suitable thickness (20 nm or less) is then epitaxially grown. Selective etching of the SiGe layer leaves an empty space (air-gap) underneath the film. The suspended Si membrane can be used to fabricate gate-all-around (GAA) transistors. Alternatively, the air-gap can be filled with a dielectric in order to form a localized SOI structure, integrated in the bulk-Si wafer [89,91].

4.4 Advanced Wafer Engineering

The International Technology Roadmap for Semiconductors points out that the performance of Si devices should be improving by approximately 17% per year. For many years now such improvements were being achieved primarily by shrinking the device dimensions, i.e., device scaling and advances in lithography provided the needed performance. More recently, it has been realized that just shrinking the device dimensions is no longer sufficient.

New or improved materials and novel configurations of existing materials have become essential. "New" materials and modification of existing materials are now pervasive in IC manufacturing. They include Cu interconnects as a replacement for Al, low-k dielectric to separate the layers of interconnects, development of high-k dielectric to replace the traditional SiO_2 gate dielectric. In addition to these device-scale improvements, modifications to the starting wafers can also greatly enhance performance of the circuits that are built on them [92].

Below, we describe some of such wafer-scale modifications. They range from extremely simple solutions, such as rotation of the wafer crystalline plane, to much more complex in which charge carrier mobilities are enhanced by modifying basic properties of silicon, i.e., its lattice parameters. In almost all of these wafer-scale solutions, the ability to bond wafers, and in particular a transfer of layers from one substrate to another that is enabled by Smart Cut, plays a key role.

4.4.1 Crystal Orientations

4.4.1.1 A Rotation within the (100) Plane

Silicon-on-insulator wafers produced by layer transfer add a new degree of freedom to the fabrication process, namely the possibility of choosing different crystal orientations for the thin active layer and the handle wafer below it. Traditionally, Si wafers were made with the (100) crystalline orientation of the surface and a notch (or a flat in smaller diameter wafers) that defined the $<110>$ direction in the wafer plane. The (100) surface was chosen in the early days of MOS technology as it yielded the lowest density of interface states, D_{it}. With advanced surface preparation techniques available today, this selection is less significant. The notch (flat) location defines the alignment of the edges of rectangular silicon die with the (110) preferred cleavage planes of Si. This alignment was essential when wafers were divided into die by

scribing and breaking, and it is still of some importance today when wafers are cut with a saw, since it helps to maintain smooth edges of the die.

Currently, when every enhancement of charge mobility is important, the traditional configuration with transistor channels in the (100) plane and the current flow in the <110> direction is less than ideal. Specific shapes of conduction and valence bands in silicon lead to different optimum configurations for the mobilities of electrons and holes [93]. For the NMOS, electrons have the highest mobility in the conventional (100) plane, and both <100> and <110> directions of the current flow are approximately equivalent [94,95]. But for PMOS, there is about 16% mobility advantage when the channel in the (100) wafer is aligned with the <100> direction [94]. This can be easily implemented in SOI by rotating the thin film by 45 degrees around the axis normal to the wafer plane, so that the <110> in the handle coincides with the <100> direction of the active device film (See Figure 4.11) [95].

4.4.1.2 Other Crystalline Planes

Because of the difference in the shape of electronic bands for electrons and holes, the crystalline plane that provides the best performance for NMOS devices, namely the (100), is far from optimum for the PMOS. PMOS performance can be greatly enhanced simply by choosing wafers with the (110) surface plane. Hole mobility is then almost doubled in comparison to the (100) configuration, and it is true even for extremely thin (110) oriented films, of the order of 6 nm [96]. Unfortunately, this comes at a significant penalty to the NMOSFET performance, as shown in the data of Figure 4.12 [97].

The question then arises what is the best way to independently optimize crystalline orientations for NMOS and PMOS. One option relies on using conventional wafers and varying the device architecture [98], another option involves wafers that are engineered to provide access to both crystalline orientations [97,99].

Nonplanar transistors such as FinFETs are described in Section 4.10 of this chapter. It suffices to say here that in such structures the transistor channel plane is perpendicular to the wafer plane. Doris et al. [98], have proposed utilizing a mix of planar transistors for NMOS and FinFETS for PMOS in order to get both the (100) plane for the former and the (110) plane for the latter, assuming that the current flow direction in the PMOS is parallel to <110>. The main problem with this scheme is that the circuit architecture becomes complex, and integration of planar and non-planar devices on the same wafer is not trivial.

The second option for simultaneous use of two different crystalline orientations moves the burden from device fabrication to the wafer level. Hybrid orientation technology (HOT) is based on forming SOI wafers in which crystalline orientation of the Si film is different than that of the handle wafer [97,99]. It should be clear that only technologies based on wafer bonding make such structures possible.

HOT SOI wafers are typically made by Smart Cut, so that the Si film is (100) and the handle is (110) or vice versa. In the first case, NMOS devices are made in SOI, while regions that will be dedicated to PMOS

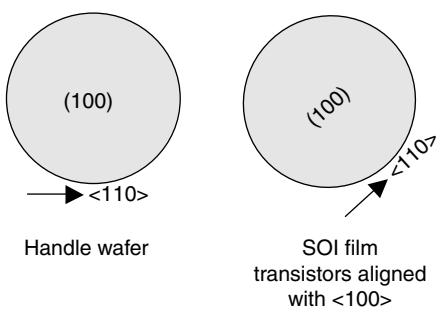


FIGURE 4.11 The concept of a 45-degree rotation of the Si film on insulator.

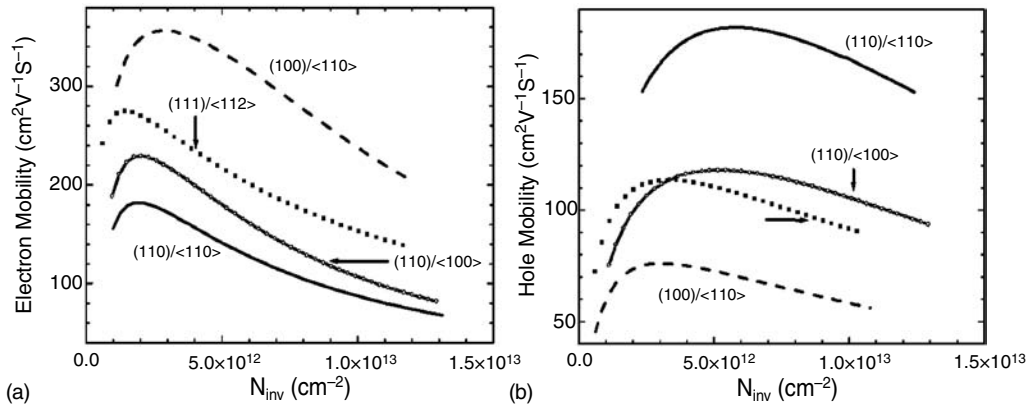


FIGURE 4.12 Electron and hole mobility for various crystal orientations. (From Yang , M., Ieong, M., Shi, L., Chan, K., Chan, V., Chou, A., Gusev, E., et al., *IEDM*, 2003.)

are etched down to the handle and refilled by selective Si epitaxy in order to obtain a reasonably plane wafer surface. The final structure is shown schematically in Figure 4.13.

4.4.2 Strained Silicon-on-Insulator

Strained silicon-on-insulator (strained SOI) refers to a composite wafer substrate that combines two high performance technological solutions: strained silicon and thin SOI [100,101]. The advantage of strained silicon lies in its electrical properties. The crystalline lattice of the top, electrically active layer of silicon is distorted in a way that enhances the mobility of electric charges, thus leading to improved transistor performance.

It has been known for a long time that electron and hole mobilities are affected by straining the crystalline lattice—in other words by changing the spacing between atoms. Any strain that is not hydrostatic, (i.e., uniform in all directions), introduces an asymmetry that removes degeneracy in the conduction and valence bands [102,103].

Already in 1982 Manasevit [104] speculated that higher electron mobility that he observed in SiGe/Si superlattices might be caused by a relatively large "lattice and thermal mismatch." During 1980s several research groups, at AT&T Bell Labs, Daimler Benz, and IBM among others, worked on optimizing growth of SiGe on Si and on understanding the physical phenomena associated with lattice strain and band offsets [102]. The work was initially stimulated by the concepts of modulation-doped superlattices built in III–V semiconductors [105] and by the requirements for Si-based heterojunction bipolar transistors or HBTs [106].

A major impetus to expand material activities for strained Si formation was added by FET device papers, starting with a 1992 article from Stanford by Welser et al. [107] that clearly demonstrated a

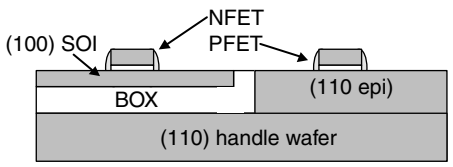


FIGURE 4.13 Hybrid Orientation Technology (HOT) concept.

substantial mobility enhancement in nMOS transistors with strained Si in the surface channel. Welser's paper excited the imagination of material scientists who have redoubled their efforts to design optimized layers, and of device engineers who aimed at the highest device performance.

When the cubic symmetry of a silicon lattice (face-centered diamond lattice) is broken, the energy degeneracy is lifted in both the conduction and the valence bands. For a biaxial tensile strain on a (100) wafer (the film is uniformly strained in the wafer plane), six equivalent energy valleys in the conduction band split into a set of four that are in the plane and have a higher energy than the two valleys that are out of plane, as shown in Figure 4.14. The net result for electrons is as follows: the interband scattering between equivalent valleys is reduced, and the in-plane effective mass of electrons on the lower energy level Δ_2 is smaller. Both effects contribute to a higher effective mobility [108,109].

For holes, the degeneracy at $k=0$ in the inverse lattice space is also lifted, and the band that is known as LH (for light holes) in the absence of tensile strain, moves up with respect to the heavy holes (HH) band. The band shift is, however, significantly smaller than in the case of electrons, and the bands become highly anisotropic. For moderate amounts of strain, there is little or no hole mobility enhancement.

When a thin layer of single crystalline semiconductor is grown on a semiconductor with the same lattice symmetry, but a slightly different lattice constant, the growing film conforms to the atomic spacing of the layer below. As the film thickness increases beyond a critical value, the energy associated with the homogenous strain increases to a point where it becomes energetically favorable to form misfit dislocations and the strain is then released [110]. The most common technique for forming Si films that are biaxially strained is by epitaxial growth on top of a single crystalline Si_xGe_{1-x} alloy layer. The lattice constant of pure Ge is 4.2% greater than that of Si. Since Ge and Si are miscible in any concentration, it is possible to adjust strain in a very thin layer to any value between 0 and 4.2%. However, because of the segregation effects, it is not practical to grow from the melt large SiGe crystal ingots of uniform composition that could be cut into wafers. Instead, "virtual" SiGe wafers are employed. Such a wafer has, as its top layer, relaxed monocrystalline SiGe with a lattice spacing that is defined by the ratio of Si to Ge.

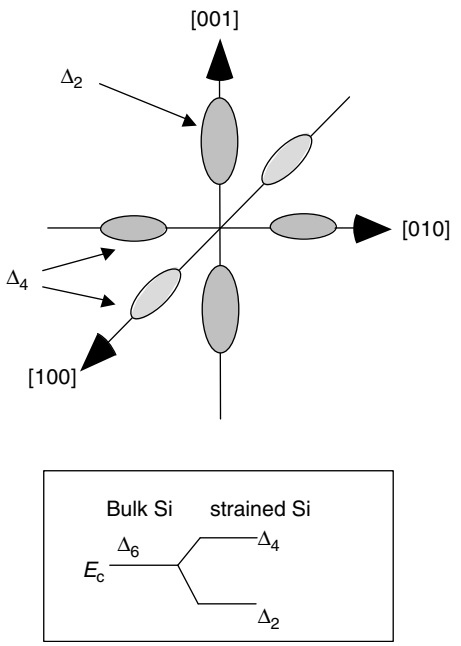


FIGURE 4.14 Splitting of degenerate conduction bands in silicon under the influence of biaxial strain.

Although it is possible to produce bulk wafers with a strained Si layer at the top, "on-insulator" structures are particularly suitable for producing and utilizing strained layers. Strained SOI, typically denoted as sSOI (and sometimes for strained Si directly on insulator; SSDOI), consists of a layer of Si that has been strained by epitaxial growth on a layer of relaxed SiGe, and then transferred to a new substrate.

To get to that point, a good quality layer of relaxed SiGe with desired Ge content has to be grown first. The technology for obtaining such SiGe layers usually requires a series of steps, often involving the epitaxial growth of a thick alloy film with the Ge composition gradually rising from 0% to the final value of about 20%. By setting the growth conditions properly, misfit dislocations that accommodate stress resulting from the lattice mismatch are confined within this thick "graded buffer" layer and do not penetrate into the relaxed SiGe film of uniform composition that is formed on top of it.

An alternative to the thick graded buffer is a thin strain-relaxed buffer, where the relaxation is facilitated by intentional introduction of nucleation sites for formation of misfit dislocations. These nuclei can be formed by very low-temperature epitaxial deposition of Si or SiGe [111], by C doping [112], or by ion implantation [113].

Smart Cut™ technology makes it possible to transfer a thin layer of relaxed SiGe with an even thinner layer of strained Si grown on top of it from a substrate on which it was initially grown to a new handle wafer [100,101]. After hydrogen implantation and bonding to another wafer, the original template can be removed without danger of relaxing the Si film. Figure 4.15 demonstrates transfer of a Si film and some of the underlying SiGe to a new substrate. Removal of SiGe by selective etching completes the process.

Another flavor of the same process results in Si on SiGe on insulator, and it is usually known as SGOI. In contrast to the previously described process, here just a layer of SiGe that is capped with oxide is transferred to another wafer, and this is followed by epitaxial growth of strained Si on top of the structure that consists of SiGe on BOX on a handle wafer. The drawback of SGOI is that the total thickness of the transistor body is relatively large, thus amplifying the short-channel effects. SGOI process was developed first and it was initially believed that sSOI was either not possible at all (compliant substrate) or limited to very thin films that could only be used for FD device applications.

In Figure 4.16, Raman spectra are shown for a thin strained Si film that is still on a donor wafer (starting material) and after transfer to a handle wafer (strained SOI or sSOI) [114]. The key point is that there is no frequency shift between the two peaks, indicating that the strain is preserved in the

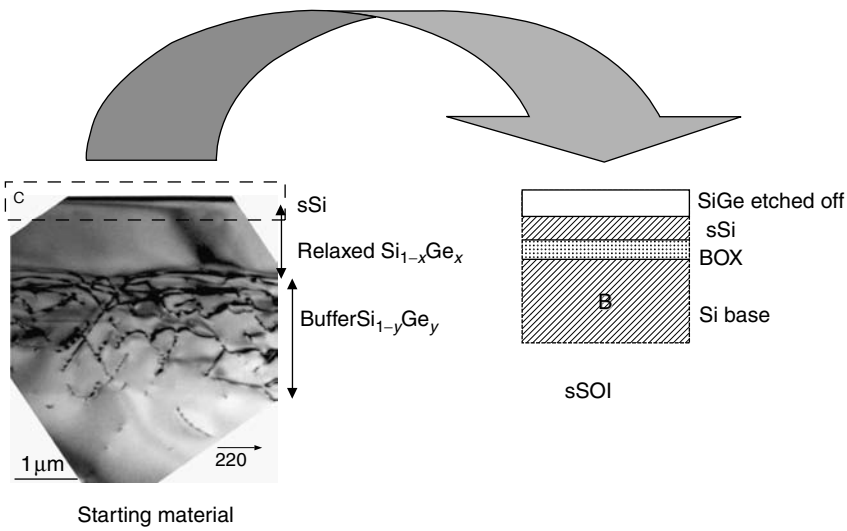


FIGURE 4.15 A layer-transfer process that results in an sSOI structure.

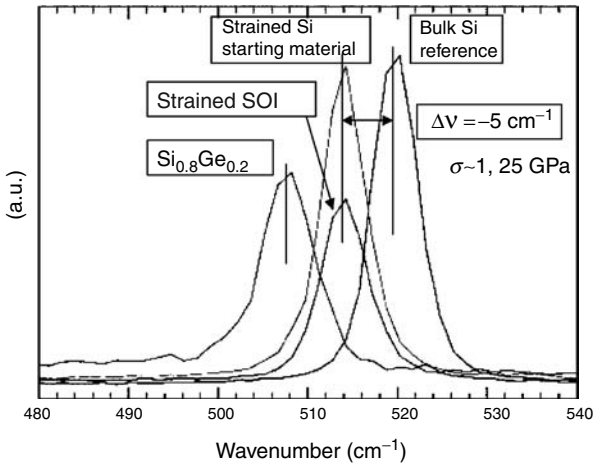


FIGURE 4.16 UV Raman data that compare the degree of strain in a silicon layer that is still on a donor wafer with that after the layer transfer and etching away of the SiGe film that is required to form sSOI.

Si film on oxide. Furnace annealing experiments with 15 min at any given peak temperature proved that there is no strain relaxation in sSOI at least up to 1050°C.

In addition to the layer transfer methods, SGOI (but not sSOI) can also be formed by two other methods. "Ge condensation" method pioneered by Toshiba, consists of depositing an epitaxial layer of SiGe on top of an SOI structure, sometimes followed by a capping layer of pure Si [115]. The subsequent thermal oxidation consumes Si in the SiGe alloy, while plowing Ge deeper into the film, thus increasing Ge concentration in the SiGe below the surface oxide. At the same time, the BOX layer prevents outdifusion of Ge into the substrate. Essentially all the Ge is trapped between a fixed BOX layer and the advancing surface oxide. By proper choices of temperature and ambient, SiGe with very high Ge content can be produced, and in an extreme case all of Si can be incorporated in the oxide, leaving a layer of pure Ge [116]. This has been demonstrated, but is not likely to become a practical method for making Ge-on-insulator (GeOI).

Another way of forming SGOI is based on the SIMOX concept. SiGe is epitaxially grown on Si and then it is implanted with a high dose of oxygen [117]. High temperature thermal annealing leads to the formation of a continuous BOX with SiGe on both sides of it. Since SIMOX technology requires post-implant annealing at $T > 1300$°C, only SiGe with a small Ge content is possible. At higher Ge concentrations, the SiGe would melt.

In both the condensation and the SIMOX methods, the thermal anneal conditions need to be chosen in such a way that the SiGe film be relaxed. Once this is accomplished, a strained Si layer is grown epitaxially on top.

It is also possible to make multilayer structures where SiGe layers, either a single one or multiple ones of varying composition, are intentionally strained [118]. Such structures may offer even greater enhancements in device performance than possible with simple strained Si layers.

4.4.3 Ge-on-Insulator and Other on-Insulator Substrates

When discussing silicon-based advanced engineered substrates, we should also include some important heterostructures in which other semiconductor materials are included. Ge-on-insulator (GeOI) has emerged in the recent years as a potential candidate for high performance electronics. CMOS in Ge could possibly provide higher switching speed and easier integration with high-k gate dielectrics [119,120]. Bulk Ge wafers are not a practical solution because of a limited global supply of germanium, and because

of its high specific mass, poor thermal conductivity, and mechanical fragility. Transferring a thin layer of single crystalline Ge to a Si handle wafer in order to form GeOI structures is the preferred solution [121,122].

Other important materials that can be transferred to Si-based or to non-silicon substrates include SiC, GaN, and other compound semiconductors. Since these materials are beyond the scope of this chapter, the reader is directed to a review by Di Cioccio et al. for further information [123].

4.5 Physical Characterization of SOI Wafers

Characterization of SOI wafers is much more involved than for bulk Si. The reasons are obvious—there is a lot more to measure in SOI wafers. In addition to all the standard properties associated with single crystalline Si, there are two additional layers, the BOX, and the device layer of monocrystalline Si. Film thicknesses are usually a key parameter to be established and controlled. Physical and electrical properties of two interfaces between the BOX and the Si must be tested, as well as structural defects that are either unique to SOI (such as the "HF" defects and pipes in the oxide) or more prevalent in SOI because of all the additional processing. Stress (and the associated strain), that is either introduced intentionally (see Section 4.4.2) or is the result of bonding thermally mismatched materials, needs to be evaluated. Finally, some standard inspection procedures for bulk wafers should be modified when dealing with SOI substrates because of their different optical signature. In the following subsections, physical characterization of SOI wafers is reviewed. The next major section discusses electrical characterization.

4.5.1 Si and BOX Thickness Measurements

Film thickness is measured almost exclusively by optical methods. Such measurements are nondestructive and make possible rapid evaluation of multiple points across the wafer. Reflectometry and ellipsometry are the approaches that are used both in a research setting and on a production floor. In an industrial setting, highly automated tools allow rapid mapping of wafer batches, with very large number of inspection points per wafer [124].

In spectroscopic reflectometry, a broad spectrum of collimated light impinges upon the surface of the wafer and the intensity of the reflected light as a function of wavelength is fitted to a model based on known values of refractive indices of the layers. The spectrum typically consists of a series of peaks and valleys that correspond to constructive and destructive interference of the incident light with the light that is reflected from multiple interfaces. When a broad collimated beam of light floods the entire wafer surface, a high resolution CCD array can provide simultaneously tens of thousands measurement points.

This configuration provides the most rapid mapping of entire wafers. Another approach to reflectometry is to use monochromatic light and vary the angle of incidence. This also results in a series of interference peaks that are used in fitting to a model.

Single wavelength ellipsometry allows deducing the film thickness values for known films. Spectroscopic ellipsometry provides enough measurement parameters to determine both the film thicknesses and their optical constants. Spectroscopic ellipsometry provides the most complete and accurate data, but at the cost of slow data acquisition. It is often used to calibrate more rapid reflectometry tools. For Si film thicknesses lesser than 300 Å, the accuracy of reflectometry tends to suffer, and ellipsometry may be necessary.

Although Si and BOX thickness values are obtained at the same time, the numerical values for the SiO_2, with its much lower refractive index as compared to Si, are not that precise. Sometimes, if high accuracy is needed, a destructive method is used in which the overlayer of Si is selectively etched away and the remaining oxide measured afterwards.

In the early SIMOX wafers, before greater than 1300°C annealing was incorporated into the fabrication process, there was no sharp interface between Si overlayer and the BOX. Even for the more recent generation of "standard dose" SIMOX, small Si precipitates were quite common inside the BOX. These features necessitated corrections to the curve-fitting models in order to obtain reasonable film thickness

values. Currently, all commercial SOI wafers, regardless of the fabrication method, have well-defined interfaces so these issues have subsided. However, there are new challenges ahead as strained Si layers are becoming more common. Even for the simplest structures, such as sSOI, the tabulated optical constants of silicon are no longer correct. Since strain modifies the refractive index, ellipsometric measurements of thickness and strain become coupled, and it is not always easy to partition them. In the case of reflectometry, the precise value of strain (or an independent strain measurement) is necessary before film thickness can be determined precisely. In more complex structures, such as SGOI with a layer of strained Si on top of a layer of relaxed SiGe alloy, the complexity of thickness measurement greatly increases.

4.5.2 Surface Roughness

Surface roughness is measured by atomic force microscopy or AFM, typically in a 1×1 or 2×2 μm^2 area and then also in 10×10 μm^2 area. Roughness needs to be in the 1–3 Å range to satisfy the requirements of CMOS devices. Nanotopography on a variety of spatial scales is of interest, but some ranges are difficult to measure.

4.5.3 Structural Defects

In the early days of SOI technology, very high defect densities were the predominant problem. Currently, SOI technology is mature and defects are well controlled. Since SOI wafers include two layers on top of the handle and three interfaces (handle/BOX, BOX/Si, top Si surface), potentially each of these regions could harbor defects, as shown in Figure 4.17.

Defects can be unique to the Si film and the BOX, or located at one of the interfaces. Si film may contain inclusions (SiO_2 or metallic particles), dislocations, stacking faults, divots, or hillocks on the surface. From the operational point of view, there is a category of defects known as HF defects. These are typically associated with pinholes or SiO_2 inclusions in the top Si film that permit HF solution (when the SOI sample is immersed in a etch bath) to penetrate from the surface into the BOX, thus etching out large voids that can be easily detected by optical inspection [124, 125].

Voids and Si pipes in the buried SiO_2 used to cause shorts between the Si film and the substrate. These can be detected by electrical means or by copper plating of a test SOI wafer.

Threading dislocations were the early benchmark of SOI quality, in the days when the numbers were of the order of 10^{10} cm^{-2}. Now they are more likely to be about 100 cm^{-2}.

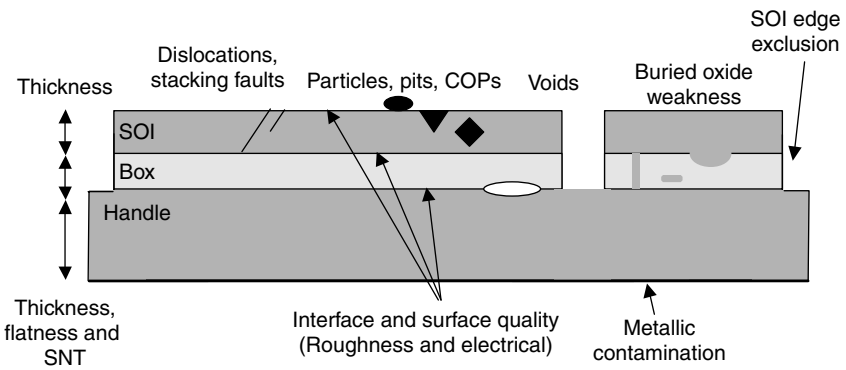


FIGURE 4.17 Drawing of potential defects in SOI layers.

4.5.4 Stress (Strain) Measurements by Raman Spectroscopy

Strain can be measured by x-ray diffraction, since it directly determines spacing between atomic planes. Such measurements can be very precise, but they are slow, complex, and often destructive. And typically the lateral spatial resolution of x-rays is low.

Raman spectroscopy has emerged as a more convenient and non-destructive test of stress in thin films. Since Young's modulus for Si is well known, values of strain are readily obtained from the measured stress. A tightly focused laser beam incident on the sample surface can provide submicron lateral spatial resolution. A judicious choice of the wavelength of light allows controlling the depth of the probed region. Since ultra-violet light penetrates only about 10–20 nm into Si, it is possible to measure stress in very thin SOI layers [126]. UV excitation energy that is close to the direct band gap of Si at 3.4 eV provides a very strong signal that is due to resonant Raman scattering [127,128]. An argon-ion laser operating at the wavelength of 363.8 nm takes full advantage of the resonant Raman signal—beam heating of the sample is reduced and the rate of data acquisition is improved.

4.5.5 Inspection for Particles and Other Defects

Wafer inspection for particles and defects is a necessary step for ensuring the quality of wafers shipped to chipmakers. The multilayer structure of SOI wafers makes it more difficult to obtain acceptable signal/noise ratio from optical scattering tools that are predominantly used for in-line nondestructive inspection. Wafer reflectivity for a given wavelength of laser light strongly depends on the specifics of the SOI stack thickness—interference effects caused by reflections from multiple co-planar interfaces can enhance or suppress the strength of the signal. Therefore, special care is needed when standard particle detection tools are used, and sensitivity to small particles can be compromised.

The problems mentioned above, all stem from the fact that for the conventional choice of lasers, the light is not absorbed strongly enough in the multilayer structure. Some of the more recently developed particle detection tools use short wavelength UV light that is fully absorbed in about 100 nm of silicon, thus eliminating any interference effects from the buried interfaces [129].

4.6 Electrical Characterization

The characterization of SOI structures is hampered by typical problems: thin and FD films, stacked interfaces, relatively thick BOX, etc. This is why the conventional methods are not all applicable to thin films, whilst novel techniques can be implemented.

4.6.1 Wafer Characterization: Ψ-MOSFET

The pseudo-MOS transistor (Ψ-MOSFET) takes advantage of the upside-down MOS structure of SOI materials (Figure 4.18). The Si substrate (gate terminal) is biased to induce a conduction channel (inversion or accumulation, electrons or holes) at the interface. The BOX is the gate oxide and the Si film represents the transistor body. Probes with adjustable pressure are placed on the film to form source and drain point contacts [130]. The measurement is feasible on the whole wafer or, preferably, on etch-defined Si islands.

In strong inversion/accumulation and the ohmic region, the $I_D(V_G)$ characteristics (Figure 4.18) follow the MOSFET-like model:

$$I_{DS} = f_g C_{ox} V_D \frac{\mu}{1 + \theta_1 (V_G - V_{T,FB})} (V_G - V_{T,FB}) \tag{4.1}$$

The form factor f_g accounts for the non-parallel distribution of current lines and can be calibrated by comparison with 4-point probe measurements. In general, $f_g \approx 0.75$, but tends to decrease in films thinner than 30 nm [131].

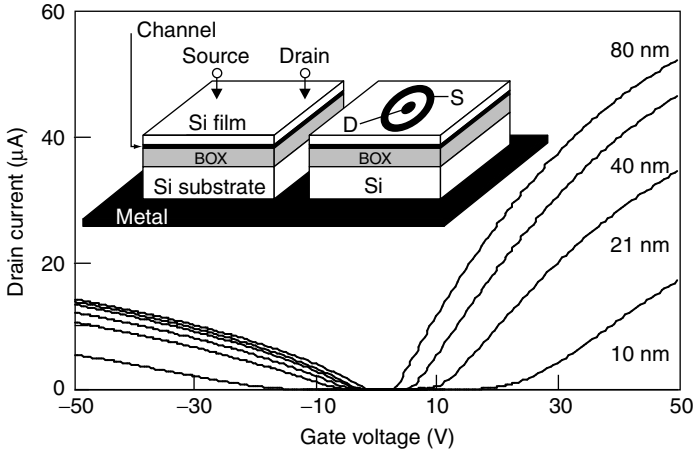


FIGURE 4.18 Configuration of the pseudo-MOS and Hg-FET transistors and typical $I_D(V_G)$ characteristics for electron and hole channels in ultra-thin film, state-of-the-art Unibond materials.

The Ψ-MOSFET can also be operated in circular configuration by using mercury probes (Figure 4.18). Such an Hg-FET features radial current lines and a well-defined f_g factor [132]. However, the measurement is longer and requires a special surface cleaning.

The slope of $I_D/(g_m)^{0.5}$ vs. V_G curves, yields the mobility μ of electrons *and* holes, whereas the intercept with the V_G axis gives the threshold voltage V_T or the flat-band voltage V_{FB}. The trap density at the film–BOX interface is calculated from the subthreshold slope, the oxide charge density from V_{FB}, and the film doping from the difference $(V_T - V_{FB})$. The carrier lifetime can also be determined by recording the transient current after the gate was pulsed in inversion [132].

The Ψ-MOSFET has been successfully tested on a variety of SOI structures: film thickness from 4 μm down to 10 nm, BOX thickness from 30 μm (SOS substrate) down to 10 nm, strained Si, Ge films, etc. The method is systematically used for process inspection and monitoring. Equation 4.1 needs, however, to be revisited for ultra-thin films by including a second-order mobility attenuation coefficient θ_2 (see for example Equation 4.4) and accounting for volume inversion effect [133].

The Ψ-MOSFET principle (i.e., substrate biasing) can rehabilitate classical methods like Hall effect, spreading resistance, and 4-point probing, which otherwise are ineffective in FD films.

An interesting development is the combination of the Ψ-MOSFET and second harmonic generation (SHG) techniques. SHG involves shining a laser on an SOI island and measuring the second harmonic of the reflected signal. The SHG signal yields the magnitude of the electric field at the interface, which is directly correlated to the channel charge. When V_G is varied, the *optical* SHG characteristic parallels the *electrical* $I_D(V_G)$ Ψ-MOSFET curve [134].

4.6.2 Device Characterization

The properties of SOI structures are inferred from the static/dynamic characteristics of the MOS transistor that stands as the main test vehicle. Parameters like mobility, threshold voltage, swing, and lifetime are determined for the front and back channels, separately or in a coupled mode. The parameter extraction is similar in SOI and bulk-Si MOSFETs and explicit methods are described elsewhere [1].

The capacitance and conductance techniques can still be applied to silicon–insulator–silicon (SIS) or MOS capacitors. The conventional theory for bulk MOS capacitors is modified to account for the formation of depletion regions on each side of the buried oxide. The limitation of capacitance and conductance measurements comes from the full depletion of the film and the large number of parameters related to the two oxides and three interfaces.

The characterization of interface traps is made by charge pumping (CP) in body-contacted SOI transistors or gate-controlled p–i–n diodes [1]. The gate is repeatedly switched from inversion (where the minority carriers are trapped on the interface states) to accumulation (where the trapped carriers recombine with majority carriers). This recombination yields an *average* CP current, which is proportional to the trap density and frequency.

1/*f* noise in MOS transistors originates from fluctuations in the carrier number (trapping) and/or mobility. The typical variation of the noise factor, S_{ID}/I_D^2, shows a plateau in weak inversion and a marked decrease in strong inversion. The magnitude of the plateau determines the density of slow traps located in the gate oxide. In extremely small area transistors, the trapping of a single carrier becomes detectable in the time domain: a small pulse *random telegraph signal* (RTS) is superposed on the average drain current value.

The signature of CP and noise is altered by interface coupling, when the back interface goes from inversion to depletion and accumulation giving rise to parasitic signals.

4.7 Partially-Depleted SOI MOSFETs

In partially-depleted SOI MOSFETs (Figure 4.2a), a neutral region subsists which leads to so-called *floating-body effects* (FBE). Designing body contacts is a compromise solution: the transistor operation does reduce to that of a bulk-Si MOSFET, but in turn, the die size, the parasitics, and the noise increase.

4.7.1 Kink Effect

Impact ionization causes the accumulation of majority carriers in the body, which triggers a reduction in threshold voltage [135]. The kink effect denotes a sudden jump in drain current (Figure 4.19a), body potential, and noise [136].

4.7.2 Hysteresis and Latch

The body charging is especially effective in weak inversion, where the current depends exponentially on potential. As the current increases, more electrons are involved in impact ionization, leading to abnormally low values of the subthreshold swing (below 60 mV/decade) and hysteresis. For high drain voltage, the body charge can sustain the inversion channel even when the gate is turned off: the transistor latches and becomes unoperational [137].

4.7.3 Parasitic Bipolar Transistor (PBT)

The lateral bipolar transistor, hibernating in any MOSFET, is brought to life by the forward biasing of the base/emitter (body/source) junction. A large collector (drain) current adds to the regular contribution of the MOS channel. The parasitic bipolar induces a premature breakdown, in both partially and FD short-channel transistors (Figure 4.19a) [138]. The PBT action can be alleviated by modifying the material properties in the channel to reduce carrier lifetime ("lifetime killing" techniques) and by source engineering: LDD to reduce the impact ionization rate, source silicidation, etc. [139].

4.7.4 Transient and History Effects

Drain current transients are observed when the floating body potential is forced out of equilibrium. The current can temporarily be higher (overshoot) or lower (undershoot) than the equilibrium value (Figure 4.19b). The transient current variation reflects the difference between the final and initial body charges, and the time constant depends on the generation–recombination process and leakage through the gate and junctions.

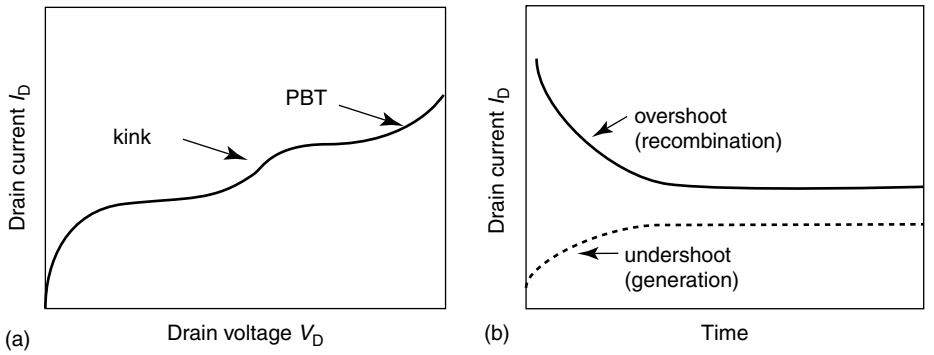


FIGURE 4.19 (a) Kink and premature breakdown in output characteristics $I_D(V_D)$, and (b) drain current transients after gate switching.

When the transistor is switched from depletion to strong inversion, the depletion region extends and the majority carriers accumulate at the bottom of the film. The body potential increases; hence the threshold voltage is lowered and induces an excess current (overshoot) (see Figure 4.19b). Subsequent carrier recombination causes the current to decrease towards the equilibrium value. The transient is processed with Zerbst-like techniques and provides the recombination carrier lifetime [1,140]. The opposite effect (deficit of majority carriers and current undershoot) is governed by carrier generation and occurs when the gate is switched from strong to weak inversion.

An experimental variant consists in biasing the front gate in inversion, whereas the back gate is pulsed into accumulation. The film potential drops, instantly lowering the inversion charge and current. The current relaxes back to equilibrium (undershoot) through carrier generation within the film and at the interfaces.

During high frequency switching of integrated circuits, the charging and discharging of the body is an iterative process that may cause "history" effects and dynamic instabilities. The switching delay of an inverter is governed by the amount of available current, which can be higher or lower than at equilibrium. The switching speed depends on the number of previous switching cycles.

4.8 Fully Depleted SOI MOSFETs

While PD SOI MOSFETs are currently used for fabricating high-performance microprocessors, FD MOSFETs are unavoidable for advanced scaling. *Full depletion* means that the depletion region covers the whole transistor body (Figure 4.2b). The FD charge does not vary with the gate voltage, which enables enhanced gate control over the inversion charge. The front- and back-surface potentials are coupled, so that the electrical characteristics of one channel vary with the bias applied to the opposite gate.

Specific $I_D(V_G)$ relations account for the characteristics of FD SOI-MOSFETs illustrated in Figure 4.20.

4.8.1 Threshold Voltage

The lateral shift of $I_D(V_{G1})$ curves (Figure 4.20a) is due to the linear decrease of the front-channel threshold voltage, $V_{T_1}^{dep}$, with increasing back-gate bias V_{G2}, between two plateaus corresponding, respectively, to accumulation $V_{T_1}^{acc}$ and inversion at the back interface [1,141]:

$$V_{T_1}^{dep} = V_{T_1}^{acc} - \frac{C_{si}C_{ox_2}(V_{G_2} - V_{G_2}^{acc})}{C_{ox_1}(C_{ox_2} + C_{si} + C_{it_2})} \tag{4.2}$$

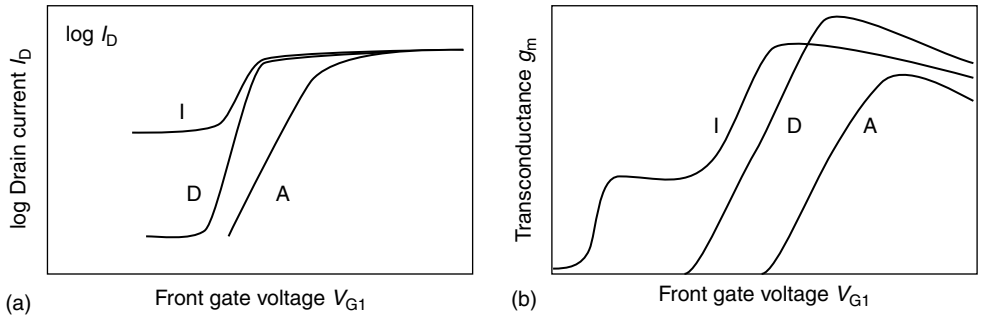


FIGURE 4.20 Generic front-channel characteristics of a FD n-channel SOI MOSFET for accumulation (A), depletion (D), and inversion (I) at the back interface. (a) $\log I_D(V_{G_1})$ in weak inversion; (b) transconductance $g_m(V_{G_1})$.

C_{si}, C_{ox}, C_{it} are the capacitances of the FD film, oxide, and interface traps. The subscripts 1 and 2 refer to the front and back channels, respectively. They can be interchanged in order to account for the reciprocal variation $V_{T_2}(V_{G_1})$.

The threshold voltage decreases in thinner films as the depletion charge is reduced. A V_T rebound is observed in ultra-thin films ($t_{si} \leq 10$ nm), due to quantum splitting of the 2D subband system [142].

4.8.2 Subthreshold Slope

The subthreshold swing, $S = dV_G/dI_D$, is minimum (close to the theoretical limit of 60 mV/decade at 300 K):

$$S_1^{dep} = 2.3\frac{kT}{q}\left(1 + \frac{C_{it_1}}{C_{ox_1}} + \alpha_1 \frac{C_{si}}{C_{ox_1}}\right) \tag{4.3}$$

for depletion at the back interface (Figure 4.20a) [143]. The interface coupling coefficient $\alpha_1 < 1$ accounts for the influence of BOX thickness and back interface traps (intrinsic or induced by radiation and hot-carrier degradation) [1].

The BOX interface defects are masked when accumulating the back channel. However, coefficient α_1 tends to unity (as in bulk-Si or PD SOI), causing an overall degradation of the swing (Figure 4.20a).

The model above is valid only when the BOX is thick enough so that substrate effects occurring underneath the BOX can be ignored. Since the capacitances of the BOX and Si substrate are connected in series, the swing for thin buried oxides also depends on the density of traps at the *third* interface: BOX-Si substrate [144].

The subthreshold slope normally improves for thinner silicon films and thicker BOX [144].

4.8.3 Transconductance

In strong inversion and the ohmic region, the front-channel drain current and transconductance are given by

$$I_D = \frac{C_{ox_1} W V_D}{L} \frac{\mu_1}{1 + \theta_1(V_{G_1} - V_{T_1}) + \theta_2(V_{G_1} - V_{T_1})^2}(V_{G_1} - V_{T_1}(V_{G_2})) \tag{4.4}$$

$$g_{m_1} = \frac{C_{ox_1} W V_D}{L} \frac{\mu_1 [1 - \theta_2 (V_{G_1} - V_{T_1})^2]}{[1 + \theta_1 (V_{G_1} - V_{T_1}) + \theta_2 (V_{G_1} - V_{T_1})^2]^2} \tag{4.5}$$

where μ_1 is the mobility of front channel carriers, and $\theta_{1,2}$ are the mobility attenuation coefficients: θ_1 accounts for the impact of series resistances and θ_2 reflects the surface roughness scattering for ultra-thin gate oxides.

The transconductance of FD MOSFETs (Figure 4.20b) is complicated by the relation $V_{T_1}(V_{G_2})$ given by Equation 4.2. The effective mobility and transconductance are maximum for depletion at the back interface, due to combined effects of reduced vertical field and series resistances [145]. The distortion of the transconductance (curve I, Figure 4.20b) originates from the premature activation of the back channel. While the front interface is still depleted, increasing V_{G_1} reduces V_{T_2} and opens the back channel before the front channel. The transconductance exhibits a characteristic plateau [145].

4.8.4 Meta-Stable Dip

Meta-stable dip (MSD) is a combination of coupling and transient mechanisms [146]. The front-channel transconductance is measured from accumulation to strong inversion (direct scan) with the back channel biased just below inversion. A surprising dip, MSD is observed for $-2.5 < V_{G_1} < -1$ V in Figure 4.21.

In "accumulation," the generation of majority carriers is not instantaneous and interface coupling occurs. As V_{G_1} increases, the *back* channel opens first, the current increases linearly, and the transconductance shows the typical plateau (as in Figure 4.20b, curve *I*). When the accumulation layer is completed and equilibrium is reached ($V_{G_1} \approx -2.5$ V), the coupling stops: the back-channel current remains constant and the transconductance falls to zero. For $V_{G_1} > -1$ V, the accumulation layer is no longer sustained by the gate and the coupling-related plateau resumes. For $V_{G_1} > 0$ V, the usual behavior of the front-channel conduction is observed.

Notice that for the reverse direction of the V_{G_1} scan, equilibrium is reached faster and there is no dip. This hysteresis may be useful for designing capacitor-less DRAMs [146].

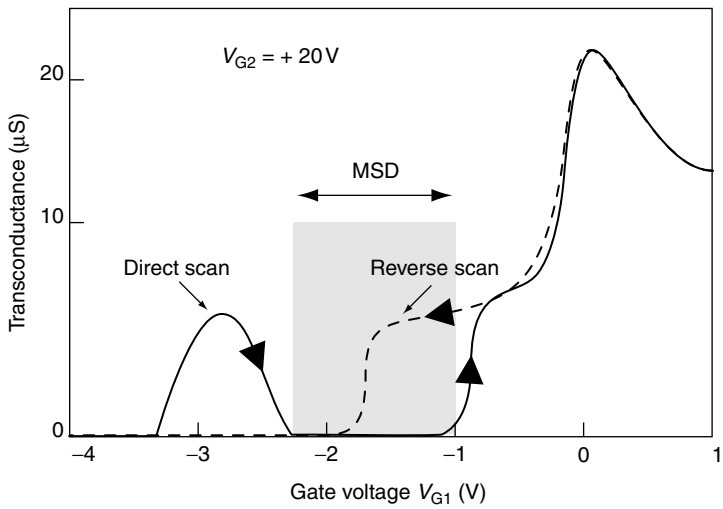


FIGURE 4.21 Front-channel transconductance vs. gate voltage for direct scan (meta-stable dip; MSD effect) and reverse scan (undoped and long channel, 80-nm thick film). The dip subsists in short-channel MOSFETs and depends on bias and time constants. (Adapted from Bawedin, M., Cristoloveanu, S., Flandre, D., and Yun, J.G., *Solid State Electron.*, 49, 1547, 2005.)

4.8.5 Volume Inversion

The simultaneous activation of front and back channels induces *volume inversion* [147]: the inversion charge covers the whole body of thin films and its maximum is located away from the interface (Figure 4.22).

For double-gate operation (Figure 4.22b), the electric field cancels in the middle of the film enabling the mobility to increase. Volume inversion subsists even in SG MOSFETs with ultra thin film (Figure 4.22a). Volume inversion, a basic mechanism in multiple-gate MOSFETs, is beneficial in terms of increased current drive and transconductance, attenuated influence of interface defects (traps, fixed charges, roughness), and reduced $1/f$ noise.

4.8.6 Transition from Partial to Full Depletion

Partial depletion occurs if the vertical depletion region w_D does not cover the whole body ($w_D < t_{si}$). This definition no longer applies to very short devices, where the lateral depletion regions of the source and drain junctions enhance the body depletion [148]. The junctions cause a lowering of the *effective* doping seen by the gate, letting the vertical depletion region to extend deeper. Therefore, the transition from PD to FD operation is not controlled by the doping/thickness ratio exclusively because the channel length contributes via a *length-to-doping* transformation: shorter the channel, lower the effective doping is.

The black region in Figure 4.23 shows the classical full-depletion domain, whereas the line between the gray and white regions indicates the revisited FD–PD boundary. In order to maintain PD operation in a 50-nm long, 100-nm thick MOSFET, a doping level (2×10^{18} cm^{-3}) more than one order of magnitude higher than in a quarter-micron device (10^{17} cm^{-3}) is needed [148].

Notice that for higher doping levels localized near the source and drain (pockets), the transition from PD to PD shows the opposite trend: shorter transistors exhibit a higher effective doping, making them more PD [149].

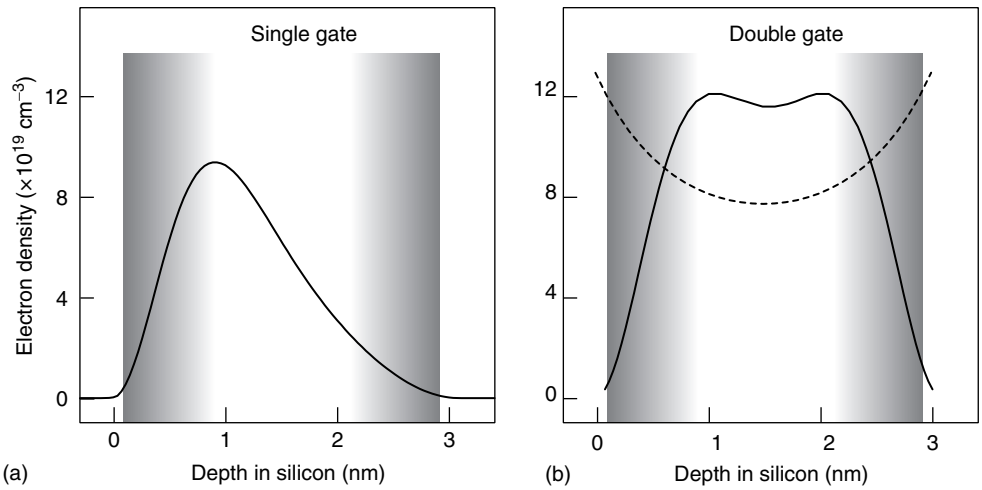


FIGURE 4.22 Minority carrier distributions in 3-nm-thick SOI MOSFETs operated in (a) single-gate mode and (b) double-gate mode with volume inversion. The dotted line reproduces non-quantum simulations, which are inappropriate for ultra-thin films. The gray sections indicate the gradual degradation of the crystal quality in the subsurface regions, where the carrier mobility may be dramatically lowered.

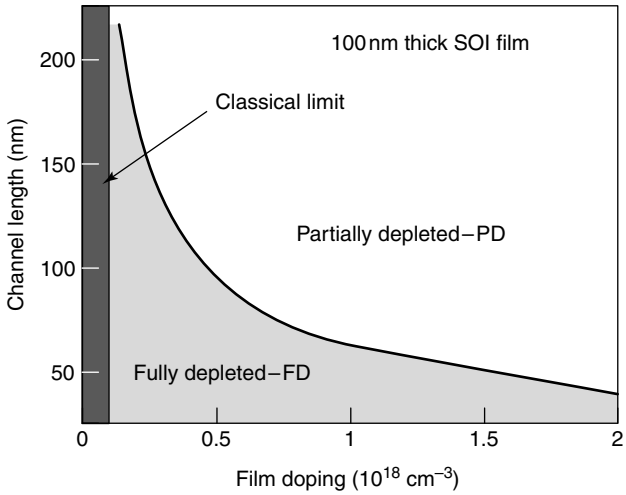


FIGURE 4.23 Transition between partial and full depletion in 100-nm thick SOI MOSFETs without pockets as a function of channel length and doping. The black region is the classical limit of full depletion in long channels. (Adapted from Allibert, F., Pretet, J., Pananakakis, G., and Cristoloveanu, S., *Appl. Phys. Lett.*, 84, 1192–1194, 2004.)

4.9 Scaling Trends

4.9.1 Short Channels

In bulk silicon MOSFETs, length scaling is performed by forming shallow junctions and increasing the body doping, which adversely affects the mobility and parasitic capacitance. The advantage of SOI is that the film thickness can entirely govern the device scaling [1,3,4].

The DIBL and charge sharing between the gate and source/drain contacts are standard short-channel effects, which occur in both bulk-Si and SOI MOSFETs. In SOI, they are better controlled by reducing the film thickness [3,4,150].

The main short-channel effect in SOI is due to the penetration of the electric field from the drain into the BOX and substrate (inset of Figure 4.24). The fringing field tends to increase the surface potential at the film–BOX interface exactly as a back-gate bias would do: this is *drain-induced virtual substrate biasing* (DIVSB) [150,151]. Since the front and back interfaces are naturally coupled, the front-channel properties degrade. In particular, the threshold voltage V_T is lowered with increasing drain bias, as in DIBL, although DIVSB is a totally distinct mechanism.

Figure 4.24 illustrates the scaling strategy in SOI. It compares V_T lowering in highly doped and undoped MOSFETs. A thick undoped film is obviously not suitable. But, for a 15-nm thick film, ΔV_T become reasonable and the doping effect disappears. It follows that an *undoped and ultra-thin* MOSFET is exceptionally robust to short-channel effects. Additional benefits are the high carrier mobility and near-ideal subthreshold slope.

Other actions against the fringing field penetration in the BOX and substrate consist in modifying the transistor architecture: thinner BOX with lower permittivity, double-gate structure, or ground plane (highly doped region or metal layer underneath the BOX) [151,152].

Electrostatic considerations show that the minimum channel length is proportional to the film thickness: $L_{min} \approx 4t_{si}$ [153]. This guiding rule means that properly scaled SOI MOSFETs with a sub-10-nm gate length will require films thinner than 3 nm [154]. As a confirmation, 2.5-nm long transistors with acceptable characteristics have been simulated for 1-nm thick SOI films [155]. On the practical side, a 6-nm long SOI MOSFET (Figure 4.25a) has already been fabricated [156].

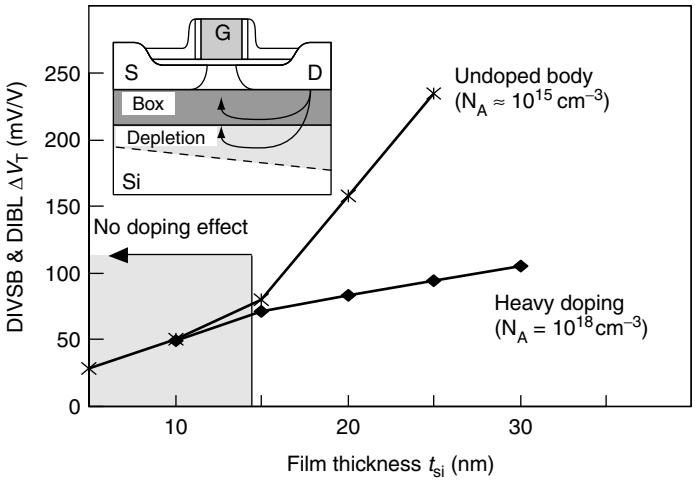


FIGURE 4.24 Threshold voltage lowering $\Delta V_T/\Delta V_D$ induced by drain-induced virtual substrate biasing (DIVSB) and drain-induced barrier lowering (DIBL) effects vs. film thickness and channel doping (channel length $L = 0.1$ μm). The doping effect is erased for films thinner than 15 nm.

Emerging technologies like strained-Si, SiGe and Ge-based MOSFETs do not compete with SOI because they *must* be SOI-like. Whatever the semiconductor, the electrostatic problems are equivalent, so the film should be ultra thin and placed on a thin dielectric.

4.9.2 Narrow Channels

A parasitic channel forms on the edges of the transistor because the gate controls differently the main channel (central region) and the edges. The impact of the sidewall channel is more prominent in narrow devices [157,158]. When the threshold voltage is lower on the edges, V_T decreases with width (reverse narrow-channel effect). Note that the parasitic conduction is normally masked by overdoping the edges such as to increase the local V_T value, but this solution fails in ultra-thin films.

Depending on the lateral isolation process (STI, LOCOS, mesa), the sidewalls may feature a different crystal orientation, additional defects, and variable thickness. In general, by reducing the width, the carrier mobility and subthreshold swing are degraded, whereas the short-channel effects show little change [157].

The FBE tend to vanish in narrow MOSFETs: the breakdown voltage, snap-back voltage and subthreshold swing in saturation are all increased, while the gain of the bipolar transistor decreases. This implies that the accumulation of majority carriers in the body is reduced when the width is scaled down. Several reasons have been documented [157]: (i) carrier lifetime degradation near the sidewalls, (ii) impurity out diffusion into the isolation oxide, which lowers the source/body potential barrier, and (iii) local thinning of the Si film edges.

In extremely narrow SOI MOSFETs, *lateral* quantum confinement leads to a remarkable increase of the threshold voltage below the 10-nm width (Figure 4.25c) [159], very much as the *vertical* confinement does in sub-10-nm thick films.

4.9.3 Channel Thickness

The film thickness defines the capability of SOI MOSFETs to withstand short-channel effects. Figure 4.25b demonstrates that film thinning down to a few monolayers can be achieved [150]. In ultra-thin FD films, the interface coupling effects (see Section 4.8) are amplified. Drawing the

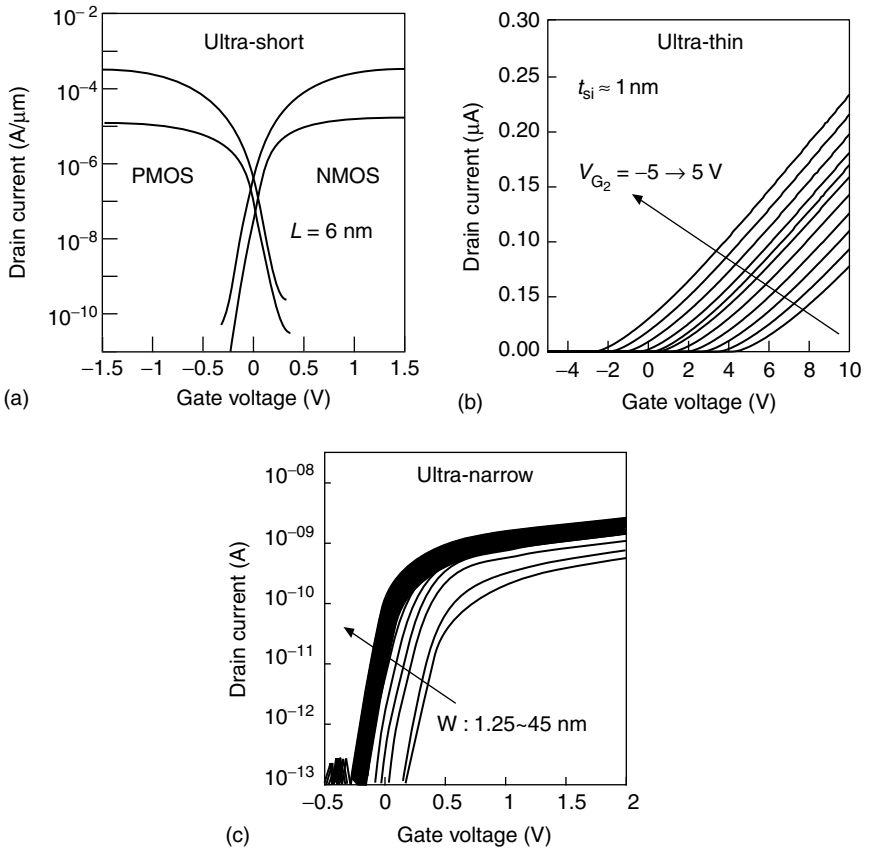


FIGURE 4.25 Drain current vs. gate voltage in ultimately small SOI MOSFETs: (a) sub-10-nm-long transistor (From Doris B., et al., *IEDM Technical Digest*, IEEE, Piscataway, NJ, 2003, 27.3.1–27.3.4.), (b) 1-nm-thick transistor (From Cristoloveanu, S., Ernst, T., Munteanu, D., and Ouisse, T., *Int. J. High Speed Electron. Syst.*, 10 (1), 217–230, 2000.), and (c) 1-nm-wide MOSFET (From Majima, H., Ishikuro, H., and Hiramoto, T., *Technical Digest IEDM'99*, 379–382, 1999.).

characteristics $V_{T_1}(V_{G_2})$ and $V_{G_1}(V_{T_2})$ on the same graph, normally results in two different curves (Figure 4.26). The intercept point labeled DG defines the unique coupling of front and back gate voltages (V_{T_1}, V_{T_2}), for which the two channels are *simultaneously* inverted [160]. These values can be used to emulate a perfectly balanced double-gate operation even when using asymmetrical MOSFETs, because they compensate for the difference in thickness between the front oxide and the BOX:

$$V_{G_2} - V_{T_2} = \frac{t_{ox2}}{t_{ox1}}(V_{G_1} - V_{T_1}) \tag{4.6}$$

In sub-10-nm-thick films, the two curves tend to coincide (dotted line in Figure 4.26) [160,161]. This implies that an arbitrary back-gate bias V_{G_2} is promoted as threshold voltage V_{T_2} as soon as the front gate is biased at threshold ($V_{G_1} = V_{T_1}$): when one channel reaches strong inversion, the opposite channel is also dragged into inversion (super-coupling). The film behaves as a quasi-rectangular well, the potential of the entire film following the signal applied to either gate. It becomes impossible to create an inversion channel facing an accumulation channel.

Since the minority carriers are no longer confined at the interface, the notion of front and back channels becomes obsolete and needs to be replaced by the concept of *volume inversion* [147].

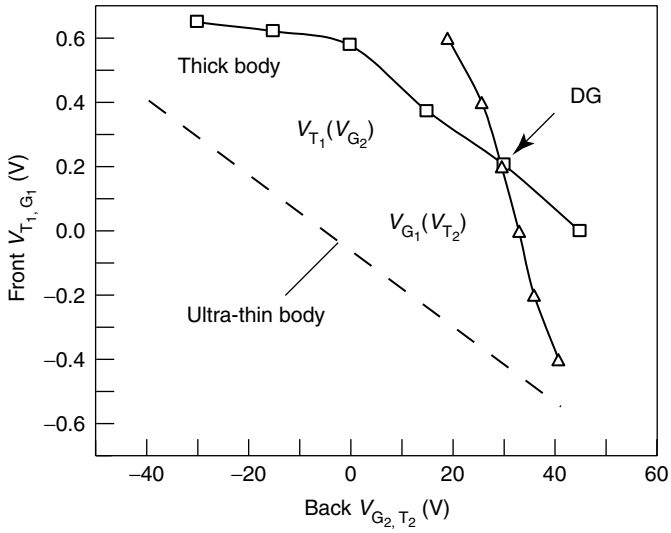


FIGURE 4.26 Front (back) channel threshold voltage vs. back (front) gate bias in a 47-nm-thick SOI MOSFET ($L=$ 10 μm, $W=10$ μm) (From Pretet, J., A. Ohata, F. Dieudonné, F. Allibert, N. Bresson, T. Matsumoto, T. Poiroux, J. Jomaah, S. Cristoloveanu. In *Silicon Nitride and Silicon Dioxide Thin Insulating Films VII*, Electrochemical Society Proceedings, Vol. PV-2003-02, 476–87. Pennington, NJ: Electrochemical Society, 2003.) Point DG indicates the appropriate bias for balanced channels in pseudo double-gate operation. The dotted line shows the superposition of the two curves (super-coupling) in ultra-thin transistors.

This implies that the conventional formulation of "front-channel mobility" should be translated into "mobility seen from the front gate."

4.9.4 Mobility Issues

In sub-10-nm-thick films, vertical quantum confinement and sub-band splitting render the carrier mobility subject to competing effects: lower effective mass and enhanced phonon scattering [162,163]. The centroid of the inversion charge is moved towards the volume of the film; hence the surface roughness scattering is also reduced [162]. Monte Carlo simulations suggest that the mobility is maximum in 3–5-nm thick films [164], but this feature still needs to be experimentally demonstrated. The opposite trend, i.e., mobility degradation in thinner films, was frequently observed.

The mobility-thickness correlation was investigated by probing different areas of a wafer, where local thickness variations subsist [165]. Figure 4.27a shows that only in shorter-channel, 10–20-nm thick transistors the mobility decreases for thinner films. This implies that the mobility degradation might be a series resistance artifact. The series resistance effect is marked in short-channel transistors and increases in thinner films.

The accurate extraction of the carrier mobility in ultra-thin SOI transistors may also be affected by [165]:

- Film thinning by sacrificial oxidation generates defects (stacking faults, dislocations, etc.) that subsequently may degrade the quality of the gate oxide interface.
- Polysilicon depletion and quantum confinement lower the effective gate capacitance. In MOSFETs with ultra-thin gate oxide, the front-channel mobility can be underestimated by more than 35%.
- The substrate acts as a back gate with infinite overlap (see Section 4.10.1) lowering the series resistance of the back channel.

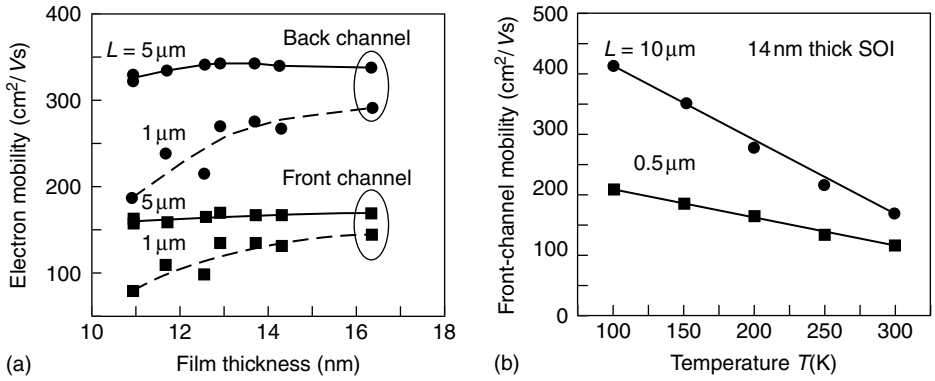


FIGURE 4.27 (a) Front and back channel electron mobility vs. film thickness in short and long SOI MOSFETs. (b) Front-channel electron mobility vs. temperature. (From Ohata, A., Cassé, M., Cristoloveanu, S., and Poiroux, T., *Proceedings of ESSDERC 2004*, IEEE, Piscataway, NJ, 2004, 109–112.)

- A fair comparison requires that the back surface potential (not the back-gate bias!) remains constant. This is not trivial because the back-channel V_T changes with thickness. Keeping the back interface accumulated is not suitable either, because of gate-induced FBE (Figure 4.29b).

The mobility behavior at low temperature may apparently look different according to the channel length (Figure 4.27b) and substrate bias. The attenuated mobility-temperature dependence in short device is actually a series resistance effect and does not reflect a prevailing Coulomb scattering.

What is the most reliable technique for mobility extraction? We have seen in Section 4.6.1 that the function $Y(V_G) = I_D/g_m^{0.5}$ is a good option. It can be extended to the case of ultra-thin oxides by including the effect of the second mobility reduction factor θ_2 as shown in Equation 4.5. The split C–V method is equally popular because it leads to the *universal* curve of mobility vs. vertical field. The mobility is determined from the ratio between the drain current and the inversion charge that is computed by integrating the C–V curve. In FD MOSFETs, however, the C–V plot measured at high frequency between the gate and source/drain terminals shows typical features (Figure 4.28a). The gate oxide capacitance is obtained in inversion but the minimum value, measured for front interface depletion, depends on the back-gate bias. Inversion at the back interface screens the effect of the BOX and substrate, so that the minimum value yields the series combination of the depleted film and gate-oxide capacitances [166]. The split C–V method fails in weak inversion, in very short transistors, and is complicated for back-channel measurements.

In this context, the *geometrical magnetoresistance* is an attractive method for determining the pure mobility even when split-CV is inapplicable [167]. A high magnetic field B is applied perpendicular to a short and wide MOSFET. Since the Hall field is short-circuited, the magnetoresistance is $R_B = R_0(1 + \mu^2 B^2)$. The carrier mobility is extracted from the linear $R_B(B^2)$ plot without needing the effective channel length (Figure 4.28b).

4.9.5 Ultra-Thin Gate Dielectrics

A special FBE takes place in MOSFETs with ultra-thin ($\leq$2 nm) gate oxide, where the body is charged by the gate tunneling current, giving rise to *Gate-Induced Floating-Body Effect* (GIFBE). The GIFBE occurs even at low drain voltage and is not related to impact ionization.

The body potential is defined by the balance between the incoming gate tunneling current (body charging) and the outgoing current (body discharging via junction leakage and carrier recombination). In PD MOSFETs, the increase in body potential directly lowers the threshold voltage [168], giving rise to

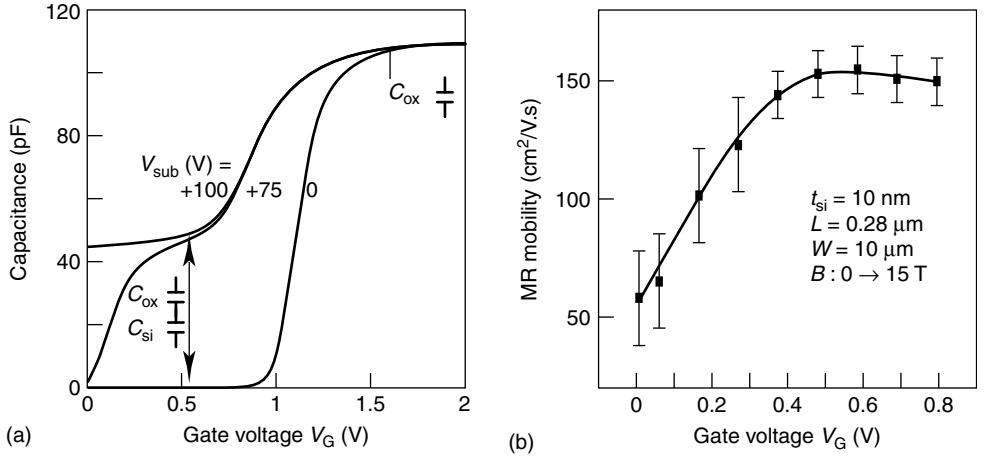


FIGURE 4.28 (a) Split-CV measurements in a FD n-channel MOSFET and (b) magnetoresistance mobility in weak and moderate inversion. (From Chaisantikulwat, W., Mouis, M., Ghibaudo, G., Gallon, C., Fenouillet-Beranger, C., Maude, D. K., Skotnicki, T., and Cristoloveanu, S., *Proceedings of ESSDERC 2005*, edited by Ghibaudo, G., Skotnicki, T., Cristoloveanu, S., Brillouet, M., IEEE, Grenoble, September. 2005.)

a "kink" in the drain current [169], a second g_m peak (Figure 4.29a), and an excess $1/f^2$ noise superimposed on the conventional $1/f$ noise [170].

The model differs for FD transistors (Figure 4.29b), where GIFBE increases the potential at the film–BOX interface. The front-channel threshold voltage is then *indirectly* lowered by interface coupling effect (Section 4.8.2) [171].

Gate-induced floating-body effect is a dimensional effect, which decreases with length because the tunneling current ($\sim LW$) is reduced while the junction leakage ($\sim t_{si}W$) is constant. Gate-induced floating-body effect also decreases in narrower MOSFETs, where the carrier lifetime and source-drain barrier are lowered near the sidewalls. For slower measurements, the second peak of the transconductance appears at a lower gate voltage. The asymmetry between gradual body charging (for increasing V_G) and body discharging (for decreasing V_G) is summarized by a hysteresis in $I_D(V_G)$ curves [160,168].

The transient drain current overshoot and undershoot (see Section 4.7.4) and history effects are drastically modified. First, the tunneling current enables a faster recovery of the equilibrium body charge. Second, the charge stored in the body can prevent the body potential to fall when the gate is turned off. While expecting an undershoot, one may observe an overshoot [160]. In general, GIFBE allows reaching the steady state more rapidly which means shorter "history effects" in digital circuits [160].

In FD MOSFETs, the second peak in transconductance is markedly amplified when the back interface is driven towards accumulation (by substrate biasing or radiation effects). An interesting consequence is shown in Figure 4.29b: the GIFBE peak gradually distorts and eventually offsets the mobility-related first peak. The mobility extracted from such a curve is meaningless because the transconductance maximum is governed by the body potential, not by the carrier mobility.

The implementation of high-k dielectrics is the best option to suppress gate tunneling current effects and enhance the transconductance. Technology improvement is still needed to reduce the interface defects that are equally detrimental in SOI and bulk MOSFETs, as compared to the near perfect Si–SiO_2 system. It is interesting to note that the direct comparison between SiO_2 and high-k is straightforward in SOI and requires a single transistor with a high-k gate stack. The front-channel characterization reflects the properties of the Si/high-k interface to be compared with those of the back channel located at the Si–SiO_2 BOX interface. A clear difference was observed in the

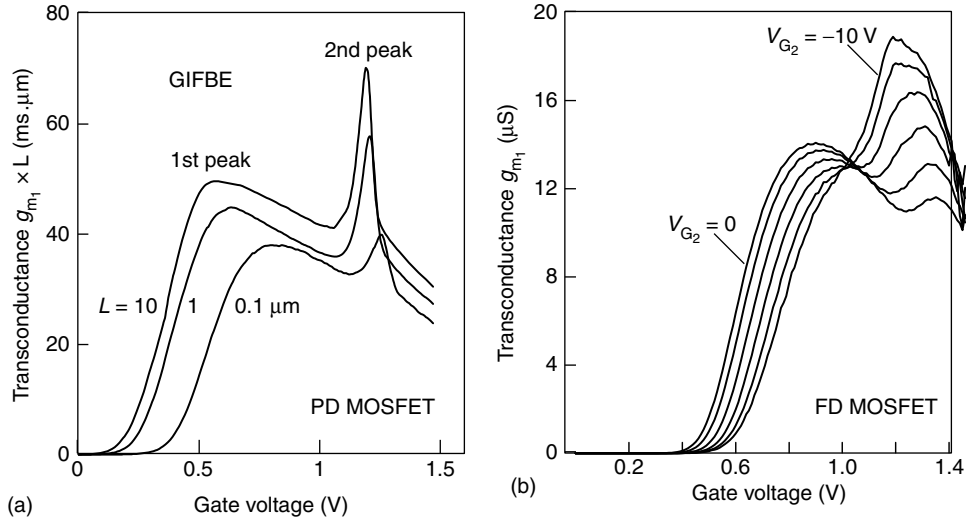


FIGURE 4.29 Second peak in transconductance (gate induced floating-body effect; GIFBE) in SOI MOSFETs: (a) short and long PD transistors and (b) FD SOI MOSFETs with variable back-gate bias ($L = W = 10$ μm, $t_{si} = 17$ nm, $V_D = 0.1$ V; From Cassé, M., Pretet, J., Cristoloveanu, S., Poiroux, T., Fenouillet-Beranger, C., Fruleux, F., Raynaud, C., and Reimbold, G., *Solid State Electron.*, 48 (7), 1243–1247, 2004.).

low-temperature mobility behavior: dominance of acoustic phonon scattering for SiO_2 and of remote Coulomb scattering for HfO_2 [172].

4.9.6 Novel BOX

The BOX needs to be optimized in order to face the problem of self-heating. In thin SOI MOSFETs, the heat path through the source/drain regions is squeezed and the body temperature rises dramatically (Figure 4.30a), lowering the mobility and threshold voltage [173]. Self-heating is primarily due to the poor thermal conductivity of the BOX, which blocks the heat dissipation into the silicon substrate. By comparison, the heat flow through the front-gate stack or source/drain terminals is less relevant. A reasonable solution is to replace the standard SiO_2 BOX with buried alumina [174,175] or other dielectrics [176] featuring an improved thermal conductivity. These are still generic SOI structures, but with the BOX other than silicon dioxide.

The thermal superiority of the novel dielectrics over SiO_2 is more visible in shorter channel MOSFETs. There is no BOX thickness effect, if the BOX is either extremely conductive (diamond, SiC) or isolating (air). BOX thinning from 400 to below 50 nm is reasonable only for Al_2O_3, quartz and SiO_2 (Figure 4.30a).

Even a modest reduction in self-heating from 100 to 50°C represents an immediate 25% gain in mobility ($\mu \sim T^{-1.5}$) for both electrons and holes. Mobility engineering can therefore be simply achieved by preventing excessive self-heating.

The change in the dielectric constant also impacts the electrostatics of the transistor by modifying the 2D distributions of the electric potential. For example, the coupling between the front and back channels (Equation 4.2) in MOSFETs with buried Al_2O_3 is three times higher than for SiO_2.

2D simulations [175,176] show that the classical charge sharing effect is marginally degraded for high-k BOX: only 25% larger V_T roll-off in 25-nm-long, alumina-BOX MOSFET [176]. The control of the fringing fields (DIVSB) is acceptable ($\Delta V_T/\Delta V_D \leq 100$ mV/V, $S \leq 100$ mV/decade) for diamond, quartz, SiO_2 and air, or modest (250 mV/V) for SiC and Al_2O_3. Quartz and diamond are best-suited dielectrics for *thick* BOX, whereas for Al_2O_3, the device architecture can benefit from a ground plane (GP).

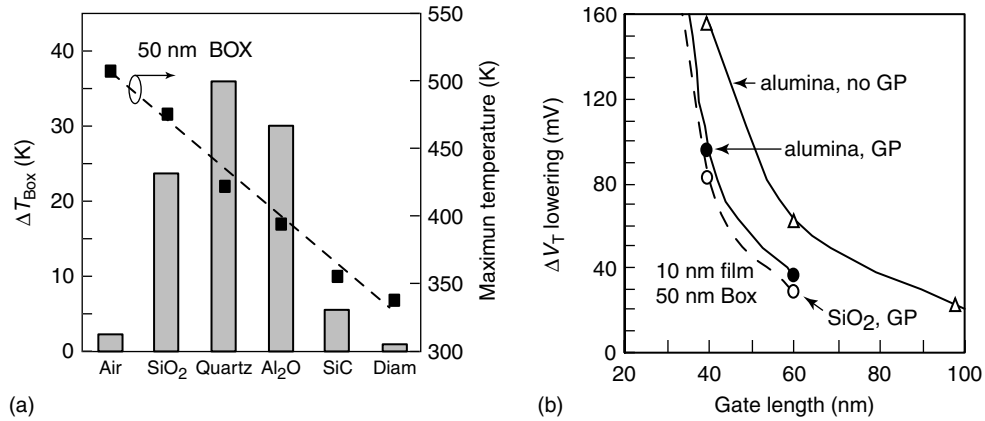


FIGURE 4.30 (a) Columns: temperature difference ΔT_{BOX} between 400 and 50-nm thick buried oxide (BOX); Line: body temperature for several BOX dielectrics (50-nm long MOSFETs, after Bresson N., S. Cristoloveanu, C. Mazure, F. Letertre, H. Iwai. *Solid State Electron.*, 49(9), 1522–1528, 2005.) (b) Threshold voltage reduction, induced by DIBL and DIVSB, vs. channel length in various FD MOSFETs (SiO_2 or Al_2O_3 BOX, ground-plane (GP) or no GP, From Oshima, K., Cristoloveanu, S., Guillaumot, B., Iwai, H., and Deleonibus, S., *Solid State Electron.*, 48, 907–917, 2004.).

Figure 4.30b depicts combined solutions: ultra-thin Si film (5–10 nm), thin BOX (50 nm), and ground plane. Without GP, DIVSB effect increases exponentially in MOSFETs shorter than 50 nm. A GP is more effective for shorter channels and alumina BOX. We conclude that the electrostatic disadvantage of alumina or high-k BOX is minor compared to the thermal advantage. It is also clear that these novel dielectrics are tremendously attractive for temperature management in high-power SOI devices.

4.10 Multiple-Gate SOI MOSFETs

4.10.1 Double-Gate MOSFETs

The DG MOSFETs are ideal devices for electrostatic integrity and ultimate scaling. Not only is the total inversion charge in DG-mode twice as large as in SG mode, but also volume inversion is enabled. The current is higher, the subthreshold swing is nearly 60 mV per decade at 300 K, and the carriers flowing in the middle of the film experience less surface scattering. The mobility [177,178], radiation hardness [2,179], and low-frequency noise are improved. Ernst et al. [150] reported an outstanding transconductance increase, by more than 200%, for 3-nm-thick DG-mode transistor.

The short-channel effects (DIBL, DIVSB, punchthrough) are reduced; hence the minimum channel length is smaller in DG than in SG transistors [180,181]. Numerical simulations including quantum effects, band-to-band tunneling, and direct source-to-drain tunneling anticipate acceptable characteristics even for DG-MOSFETs as short as 2–8 nm [155]. The recommended body thickness-to-length ratio is roughly 1/2, a condition less stringent than in SG-MOSFETs (1/4).

The real challenge is to set a manufacturable process flow. Planar process is suitable in many respects, but hardly guarantees the self-alignment of the two gates. The DG technology can be simplified if a reasonable degree of gate misalignment is tolerable. A solution is to first fabricate a longer bottom gate on top of the SOI film. The wafer is then turned upside-down and bonded to a support Si wafer. After etching the substrate and BOX of the initial SOI wafer, a shorter front gate is formed on the denuded side of the film, roughly aligned to the bottom gate. The channel length is short, defined by the source/drain implantation through the top gate [182]. Numerical simulations show that such asymmetrical DG MOSFETs should not be disregarded: the transconductance and drive current may be higher than in "ideal" symmetrical

DG transistors [183]. This is so because the longer gate contributes to volume inversion in the body and simultaneously to accumulation in the source/drain regions. This *field-effect-junction* mechanism enables the dynamic lowering of the series resistance, which compensates for the parasitic overlapping capacitance.

A different approach is to adopt a non-planar technology. In fully vertical DG MOSFETs, the source-body-drain stack as well as the current flow is perpendicular to the wafer surface. These devices are attractive because the channel length is controlled by the epitaxial growth of the body, instead of e-beam lithography. The drawbacks are the asymmetry of the source and drain terminals, and the difficulty to achieve tiny pillars with ultra small inter-gate distances.

The FinFET is a more pragmatic non-planar DG transistor with relatively easy-to-implement process. In DG-FinFETs (Figure 4.31a), the gate covers three sides of the body (fin), but the top channel is deactivated by using a thicker dielectric. The FinFET is a semi-vertical device because the current is controlled by the two vertical gates and flows horizontally along the body sidewalls.

A more advanced alternative is to etch-off the top gate and provide independent contacts to the lateral gates. The advantage of this device, called MIGFET, is that the two gates can play different logic functions, thus reducing the complexity of digital circuits [184].

The FinFET performance is very promising, but two critical scaling issues call attention: (i) the optimization of the crystal quality and orientation of the sidewalls by wet etching, and (ii) the trimming of the transistor body (inter-gate distance) in order to control the short-channel effects [185,186].

4.10.2 Triple-Gate MOSFETs

A FinFET with a continuous top gate and active top channel is named triple-gate MOSFET (TG-MOSFET). Actually, a single gate controls two lateral channels and one horizontal channel (Figure 4.31a). The gate dielectric should be equally thin on the three sides of the body in order to avoid multiple threshold voltages. The magnitude of the current can be adjusted via the fin width, which governs the contribution of the top surface channel. This advantage is debatable because volume inversion and scaling capability are lowered in wider fins.

In Triple-Gate FinFETs, the separation of the different channels is possible by using the back-gate action and the variable fin width. It was found that the carrier mobility is significantly degraded on the fin

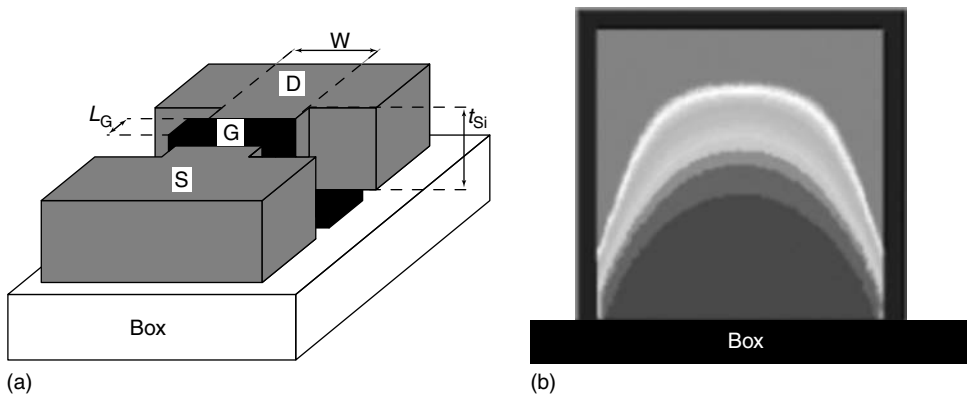


FIGURE 4.31 (a) Schematic configuration of the FinFET and triple-gate MOSFET and (b) cross-section of the minority carrier distribution in a square (20/20 nm) triple-gate device. The gate is biased in inversion ($V_{G_1} = +0.5$ V) and the substrate in accumulation ($V_{G_2} = -10$ V, From Cristoloveanu, S., Ritzenthaler, R., Ohata, A., and Faynot, O., *Int. J. High Speed Electron.*, 16(1), 9–30, 2006.)

sidewalls as compared to the top and bottom channels [187]. Process refinements [185,186] and various crystal orientations are explored to further improve FinFET performance.

It is important to understand the size effects as well as the 3D nature of the interaction between the different channels [188]: (i) *lateral* coupling between the side gates, (ii) *vertical* coupling between the top gate and the bottom gate, and (iii) *longitudinal* coupling between the drain and the body via the fringing fields (DIVSB). The coupling effects depend on the fin height, t_{si}, and width, W. A square TG FinFET ($t_{si} = W = 20$ nm, Figure 4.31b) features an inhomogeneous vertical variation of the electron concentration and surface potential on the lateral sides. The "measured" threshold voltage represents the average of the position-dependent V_T. In addition, if a large negative substrate bias is applied, the accumulation layer can block the inversion of all three electron channels.

A narrow and tall fin ($t_{si} \gg W$) exhibits two distinct regions [188]. At the bottom of the device, the carrier distribution is inhomogeneous (2D), similar to that of the square fin (Figure 4.31b). In the upper region, the potential variation along the height of the fin is negligible, except next to the top gate where corner effects may appear. The front channel and the upper regions of the lateral channels are in strong inversion, fully controlled by the gate. The substrate-to-body coupling effects are weak since the lateral coupling prevails, reducing the back-gate action. For thin and wide fins ($W \gg t_{si}$), the lateral gates are not able to control the body. Instead, the back-gate coupling is strong and modulates the front channel conduction as in FD MOSFETs.

The variation of the threshold voltage with back-gate bias depends on the aspect ratio (t_{si}/W) of the fin (Figure 4.32). For wide fins, the coupling effect $V_T(V_{G_2})$ is strong due to the classical 1D vertical coupling [141] between the front channel and the back gate. In tall fins, the lateral gates control the electrostatics, which inherently tends to suppress (for $t_{si}/W \approx 4$) the coupling to the bottom gate [188].

The geometry optimization aims to reach more current (wider transistors or multiple fins) while avoiding too much sensitivity to substrate effects. Note that even for a grounded back gate, a virtual substrate biasing can be induced by radiation, hot carrier injection, or DIVSB effects.

Further simulations show that W and t_{si} play equivalent roles. In order to suppress the short-channel effects and achieve a low-subthreshold swing, both dimensions t_{si} and W should be reasonably small or one dimension (W or t_{si}) must be *very* small.

If a very narrow body is manufacturable, the fringing fields are controlled by the lateral gates and FinFETs are suitable. The lateral gates define the back-surface potential, blocking the penetration of the

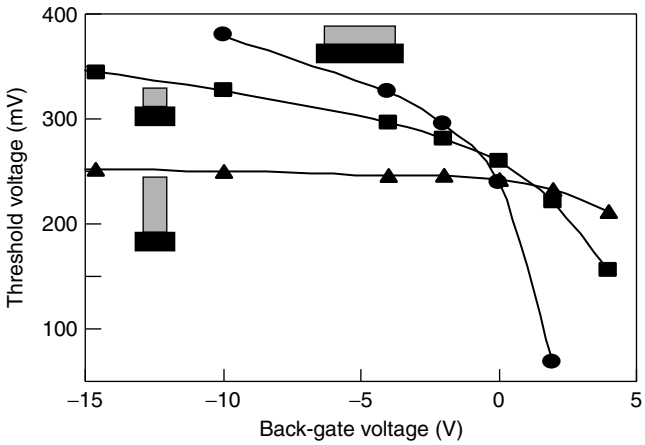


FIGURE 4.32 Threshold voltage as a function of substrate bias in triple-gate FinFETs with wide, square, and tall configurations (aspect ratios: $t_{si}/W = 20/80$, 20/20, 80/20, From Cristoloveanu, S., Ritzenthaler, R., Ohata, A., and Faynot, O., *Int. J. High Speed Electron.*, 16(1), 9–30, 2006.)

fringing field from the drain (DIVSB). This control can be enhanced by letting the lateral gates to extend vertically into the BOX (π-gate) and laterally underneath the film (Ω-gate) [2]. π-Gate and Ω-gate architectures do relax the constraint of ultra-narrow fins, but in turn require a thick-enough BOX. The question is whether these configurations can survive as the SOI materials evolve toward very thin BOX.

The opposite situation is when neither the body width nor the gate architecture enables the control of the back-surface potential. Then very thin films are needed to prevent short-channel effects, so that FD planar devices are more attractive [188].

An ultimate size effect involves the body *volume*. FinFET technology is capable of producing devices with all dimensions (thickness, width, length) in the 10-nm range (see Figure 4.25). A 10^{-18} cm^3 body raises interesting fundamental questions. For example, what doping level is induced by one impurity atom? Does the impurity position matter? Should atomistic simulations include the silicon atoms one-by-one?

4.10.3 Gate-All-Around MOSFETs

The GAA MOSFETs, invented by Colinge et al. [2] require a complex technology: (i) formation of a small size isolated Si membrane, (ii) thermal oxidation, and (iii) wrapping a homogeneous gate around. The membrane can be processed on SOI by etching part of the BOX underneath the silicon film [2], or in bulk-Si by SON technology [189]: (i) epitaxy of a sacrificial layer of SiGe, (ii) epitaxy of the thin Si film, and (iii) removal of the SiGe layer. The GAA MOSFETs can also be vertical, but the formation of a pillar with small enough diameter is very challenging.

The structure of GAA MOSFETs is conceptually simple for investigating the coupling, corner, and quantum effects. The corners are intrinsic to FinFET and GAA architectures. In each corner, the electrostatic coupling between the adjacent gates is favorable for minority carriers to accumulate [190]. The corner regions have a lower threshold voltage and turn on earlier than the main channel, which causes an increase of the leakage current in the off-state and poor subthreshold characteristics.

The corner effect increases for square bodies with high doping. Corner rounding equalizes the minority carrier distribution and suppresses the activation of the parasitic channels. A simpler solution for attenuating the corner effect is to leave the body *undoped*, while adjusting the threshold voltage with a midgap metal gate. 2D quantum simulations for GAA with very small cross-section show that quantization leads to electron repulsion from the interface and corners (Figure 4.33a) [188].

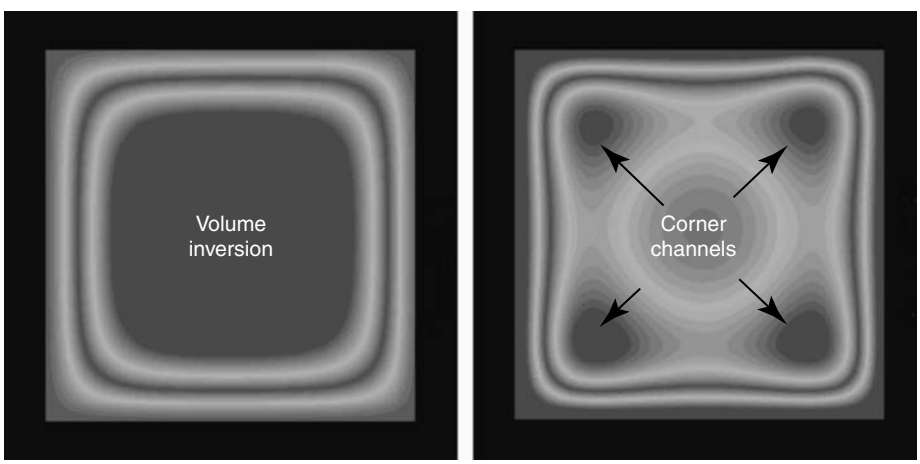


FIGURE 4.33 (a) Minority carrier distribution in a 10-nm square gate-all-around (GAA) transistor: weak inversion with volume inversion and no trace of corners, and (b) moderate inversion, where the corners start forming. (From Cristoloveanu, S., Ritzenthaler, R., Ohata, A., and Faynot, O., *Int. J. High Speed Electron.*, 16(1), 9–30, 2006.)

This repulsion opposes the electrostatic effect of corner attraction, resulting in relatively low surface and corner concentrations. Quantum repulsion and volume inversion lead to the formation of a channel in the center of the device. The subthreshold corner effect is suppressed; hence the off-state current and swing are excellent.

For higher gate bias (Figure 4.33b), the electrostatic effects are gradually taking over the quantum effects and four corner filaments are formed in strong inversion, without degrading the device performance. Nano-size and undoped fins show clear advantage for downscaling and for avoiding parasitic effects.

4.10.4 Four-Gate FET

Unlike GAA and TG-MOSFETs, the four-gate FET (G^4-FET) is a genuine four-gate transistor, operated in accumulation/depletion modes [191]. Figure 4.34a shows an inversion-mode, p-channel SOI MOSFET with two N^+ body contacts. The same device transforms into an n-channel G^4-FET when the current is carried by electrons in the perpendicular direction. The majority carriers flow between the body contacts that play the role of source and drain for the G^4-FET (Figure 4.34a). There are four *independent* gates:

- The usual front and back MOS gates govern the surface accumulation, inversion, or vertical depletion regions.
- The two lateral junctions (JFET gates) control the effective width of the body via the gradual extension of the horizontal depletion regions.

The conduction path is modulated by mixed MOS-JFET effects, from wire-like volume conduction to strongly accumulated front and/or back surface channels [192]. The G^4-FET exhibits high current and transconductance, and excellent subthreshold swing. Each gate has the capability of switching the transistor on and off. The independent action of the four gates is promising for novel applications: mixed-signal circuits, multipliers, nanoelectronic devices (quantum wires), 4-level logic functions with a reduced number of transistors, etc.

Note that the G^4-FET accommodates naturally to scaling. As the gate length of CMOS circuits goes down, the width of the G^4-FET is reduced increasing the junction gate action. However, the G^4-FET will not compete for minimum size records, it will be more suitable for innovative circuit designs.

In the depletion-all-around (DAA) mode of operation, the majority carrier channel is surrounded by depletion regions. A quantum wire can be formed (Figure 4.34b), the vertical and lateral dimensions of which are controlled by voltage bias instead of lithography. The volume-conduction channel is separated

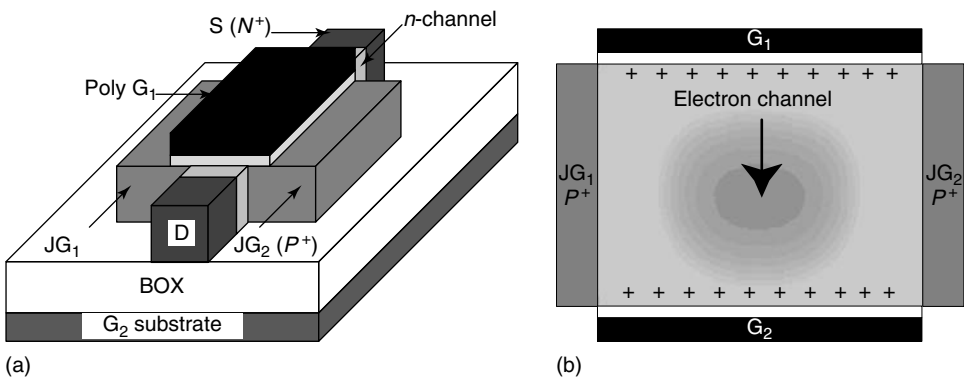


FIGURE 4.34 Basic configuration of the four-gate transistor and cross-section of carrier distribution for operation in volume mode (depletion-all-around; DAA) with inverted interfaces.

from the interfaces, first by the depletion regions and second by the inversion layers. This double shielding effect in DAA mode yields high mobility, minimum noise, and unchallenged radiation hardness capability [193]. Note that the G^4-FET is an "ambipolar" device that makes possible the independent cross-conduction of majority carriers (in the volume) and minority carriers (at the interfaces) in a single transistor.

4.11 MEMS and Photonic Applications of SOI

Classical MEMS devices are usually made in polycrystalline Si that can be readily deposited on oxidized Si wafers. Polysilicon serves well in a variety of MEMS applications and it is relatively inexpensive. However, there are many applications where deposited films of polysilicon are not the best solution. Polycrystalline Si films always have some grown-in stresses and stress gradients that are difficult to control or eliminate and these stresses can be detrimental to device operation. Polycrystalline Si surface is rough, and so is the interface with the substrate below it. Sidewalls etched in polysilicon are also rough because of the etch-rate dependence on crystal orientation—this is true for both wet and dry etching. Thickness of polysilicon films is usually limited to lesser than 3–5 μm, and MEMS applications often require thicker structures. Thermal conductivity of polysilicon is between 10 and 50% of that for single crystalline Si [194]. Torsion bars and springs made of polysilicon can exhibit fatigue and hysteresis [195] because of the change in microstructure over time.

In contrast, SOI is well suited to fabrication of various cantilevered structures with very low built-in stresses. Etching away of the BOX allows releasing any structures that are micromachined in the original SOI film. If thicker films are needed, SOI lends itself well to silicon epitaxy. A variety of MEMS devices are built commercially from SOI wafers, such as pressure gauges, gyroscopes, and also acceleration sensors for automotive air bags. Microarrays of mirrors for optical cross-connects cannot tolerate built-in stresses that would distort the mirrors–SOI wafers are ideal for making these structures.

Another emerging area for SOI is microphotonics. Just as the precise control over the flow of electrons has totally changed the world we live in, the ability to control and redirect flows of photons in microscale devices may lead to the next technological revolution. Planar photonic structures, in which layers of doped glass are deposited on flat substrate (almost always on silicon wafers since they are the most versatile) and patterned into waveguides, have been in limited use for some years. However, such structures cannot be made very small as the waveguide bend radius is limited by the difference in refractive index between the core and the cladding. For glass Δn is of the order of 0.02. On the other hand, Δn between silicon and its oxide is about two and the radius as small as 1 μm is easily accomplished. And silicon, although opaque in the visible light range, is quite transparent for the important range of near infrared wavelengths, and in particular at 1.55 μm (the best wavelength for transmission in silica optical fibers).

In order to make a silicon waveguide, a cladding of a much lower refractive index must surround the conduit of light. Silicon-on-insulator is most suitable for making such structures, with BOX preventing light loss to the substrate. Many research groups are racing to develop silicon waveguides and associated photonic devices in SOI, as evidenced by recent publications on optical modulators, silicon Raman amplifier lasers, and other applications [196–199].

It is clear that as silicon microphotonics develops, the next obvious step is to integrate it with silicon microelectronics [200]. Since same materials are used in both cases, there should be few major integration issues and we can envision some extremely high performance electronic/photonic computing and communications systems built entirely on SOI platform.

4.12 Conclusions

We have reviewed the major methods of forming SOI and SOI-like substrates, and provided relatively detailed description of the current and anticipated applications of SOI. Our main message is that SOI

devices are uniquely suitable for extending the scaling limits of conventional CMOS circuits. They also greatly improve performance, because of the inherent advantages of SOI and because of extensions of SOI to even more advanced engineered substrates, such as sSOI, hybrid orientation wafers, and various stacked materials. Strained SOI layers are of particular importance in advancing CMOS performance at the next few technology nodes.

Ultrathin Si films provide additional advantages but also extra challenges in understanding and controlling quantum confinement effects, and in terms of wafer fabrication, device processing, and characterization. Wafer engineering can also alleviate the self-heating effects, either by thinning the buried SiO_2 or by replacing it with a different dielectric that offers improved thermal conductivity.

Nano-size SOI MOSFETs provide a smooth transition from microelectronics to nanoelectronics. Thin tunneling oxides turn on remarkable gate-induced floating body effects. In nanometer-thick SOI films, the coupling effects are amplified, leading to super-coupling and interesting quantum effects.

The family of size effects in SOI devices is very rich because each dimension of the transistor plays a specific role. More importantly, a given size effect (length, width, thickness) is modulated by the other dimensions. The control of these 3D coupling effects is vital for the MOSFET scaling beyond the 10-nm channel-length barrier. What is certain is that all dimensions will be reduced concomitantly. In parallel, the transistor architecture will evolve to multiple gates, opening a wide space for new circuit topologies.

Acknowledgments

Special thanks to our colleagues from Grenoble (IMEP, SOITEC, CPMA, LETI) and elsewhere: R. Ritzenthaler, M. Bawedin, O. Faynot, J. Pretet, F. Allibert, M. Cassé, T. Poiroux, A. Ohata, K. Oshima, N. Bresson, H. Iwai, W. Chaisantikulwat, S. Deleonibus, B. Dufrene, B. Blalock, K. Akarvardar, F. Daugé, C. Gallon, A. Vandooren, J-H. Lee, P. Gentil, C. Mazuré, B. Ghyselen, I. Cayrefourcq, F. Letertre, C. Maleville, K. Bourdelle, T. Akatsu, M. Kennard, F. Brunier, O. Rayssac, and many others.

References

1. Cristoloveanu, S., and S. S. Li. *Electrical Characterization of Silicon on Insulator Materials and Devices*. Boston: Kluwer Academic Publishers, 1995.
2. Colinge, J.-P. *Silicon-on-Insulator Technology, Materials to VLSI*. 3rd ed. Boston: Kluwer Academic Publishers, 2004.
3. Celler, G. K., and S. Cristoloveanu. *J. Appl. Phys.* 93 (2003): 4955–78.
4. Frank, D. J., R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H. S. P. Wong. *IEEE Proc.* 89, no. 3 (2001): 259–88.
5. Zingg, R. P., and R. Bonne. In *Proceedings of Advanced Semiconductor Manufacturing Conference*, 2001 IEEE/SEMI, 65–6. April 2001.
6. Rudenko, T., V. Kilchytska, J.-P. Colinge, V. Dessard, and D. Flandre. *IEEE Electron. Dev. Lett.* 23 (2001): 148.
7. Colinge, J. P. In *Proceedings of International IEEE SOI Conference*, 1–4. 2004.
8. Alles, M. L., D. R. Ball, R. D. Schrimpf, D. M. Fleetwood, R. A. Reed, and B. Jun. In *Silicon-on-Insulator Technology and Devices XII*, Vol. 2005-03, edited by G. K. Celler, et al., *ECS Proceedings*, 87–98. Pennington, NJ: The Electrochemical Society, 2005.
9. Celler, G. K. *Solid State Technol.* December, (2003).
10. Xu, D.-X., A. Delage, P. Cheben, B. Lamontagne, S. Janz, and W. N. Ye. In *Silicon-on-Insulator Technology and Devices XII*, Vol. 2005-03, edited by G. K. Celler, et al., *ECS Proceedings*, 207–18. Pennington, NJ: The Electrochemical Society, 2005.
11. Cristoloveanu, S., and G. Reichert. In *High Temperature Electronic Materials, Devices and Sensors Conference Proceedings*, 1998.
12. Ohno, T., S. Matsumoto, and K. Izumi. *IEEE Trans. Electron Dev.* 40 (1993): 2074.

13. Nishimura, T., Y. Inoue, K. Sugahara, S. Kusonoki, T. Kumamoto, S. Nakagawa, M. Nakaya, Y. Horiba, and Y. Akasaka. *IEDM Tech. Dig.* (1987): 111.
14. Vogt, H. In *SOI Technology and Devices VI*, 430. Pennington: Electrochemical Society, 1994.
15. Zaslavsky, A., C. Aydin, S. Luryi, S. Cristoloveanu, D. Mariolle, D. Fraboulet, and S. Deleonibus. *Appl. Phys. Lett.* 83 (2003): 1653–5.
16. Ono, Y., Y. Takahashi, K. Yamazaki, M. Nagase, H. Namatsu, K. Kuri-hara, and K. Murase. *Technical Digest IEDM*, 367–70. Piscataway, NJ: IEEE 1999.
17. Tong, Q.-Y., and U. Gösele. *Semiconductor Wafer Bonding: Science and Technology*. New York: Wiley, 1999.
18. Plößl, A., and G. Kräuter. *Mater. Sci. Eng.* R25 (1999): 1.
19. Iyer, S. S., and A. J. Auberton-Hervé. *Silicon Wafer Bonding Technology for VLSI and MEMS Applications*. London, UK: INSPEC, 2002.
20. Alexe, M., and U. Gösele. *Wafer Bonding, Applications and Technology*. Berlin, New York: Springer, 2004.
21. Haisma, J. In *Wafer Bonding, Applications and Technology*, edited by M. Alexe, and U. Gösele, 1–60. Berlin, New York: Springer, 2004.
22. Spierings, G. A. C. M., J. Haisma, and T. M. Michielsen. *Philips J. Res.* 49 (1995): 47.
23. Kern, W. *J. Electrochem. Soc.* 137 (1990): 1887.
24. Stengl, R., T. Tan, and U. Gösele. *Jpn. J. Appl. Phys.* 28 (1989): 1735.
25. Lasky, J. B. *Appl. Phys. Lett.* 48 (1986): 78.
26. Frye, R. C., J. E. Griffith, and Y. H. Wong. US patent 4,501,060 issued Feb. 26, 1985.
27. Weldon, M. K., V. E. Marsico, Y. J. Chabal, A. Agarval, D. J. Eaglesham, J. Sapjeta, W. L. Brown, S. B. Christman, E. E. Chaban, et al. In *Proceedings of 4th Iternational Symposium on Semiconductor Wafer Bonding*, edited by U. Gösele, H. Baumgart, T. Abe, C. Hunt, and S. Iyer, 229. Pennington, NJ: The Electrochemical Society Proceedings Series, 1997.
28. Maszara, W. P., G. Goetz, A. Caviglia, and J. B. McKitterick. *J. Appl. Phys.* 64 (1988): 4943.
29. Tong, Q. Y. *Mater. Res. Soc. Symp. Proc.* 681 (2001): I1–I2.
30. Rieutord, F., B. Bataillou, and H. Moriceau. *Phys. Rev. Lett.* 94 (2005): 236101.
31. Suni, T., K. Henttinen, I. Suni, and J. Mäkinen. *J. Electrochem. Soc.* 149 (2002): G348.
32. Pasquariello, D., C. Hedlund, and K. Hjort. *J. Electrochem. Soc.* 147 (2000): 2699.
33. Wiegand, M., M. Reiche, and U. Gösele. *J. Electrochem. Soc.* 147 (2000): 2734.
34. Kaminsky, M., S. K. Das, and G. Fenske. *Appl. Phys. Lett.* 27 (1975): 521.
35. Ullmaier, H. *MRS Bull.* 22, no. 4 (1997): 14.
36. Anttila, A., J. Hirvonen, and M. Hautala. *Rad. Eff. Lett.* 57 (1980): 41.
37. Bister, M., J. Hirvonen, J. Räisänen, and A. Anttila. *Rad. Eff.* 59 (1982): 199.
38. Bruel, M. US Patent 5,374,564 issued December 20, 1994.
39. Bruel, M. *Electron. Lett.* 31 (1995): 1201.
40. Bruel, M. *Nucl. Instr., Meth. Phys. Res.* B 108 (1996): 313–9.
41. Johnson, N. M., F. A. Ponce, R. A. Street, and R. J. Nemanich. *Phys. Rev.* B 35 (1987): 4166.
42. Cerofolini, F., L. Meda, R. Balboni, F. Corni, S. Frabboni, G. Ottaviani, R. Tonini, M. Anderle, and R. Canteri. *Phys. Rev.* B 46 (1992): 2061–70.
43. Celler, G. K., A. J. Auberton-Hervé, B. Aspar, C. Lagahe-Blanchard, and C. Maleville. In *Wafer Bonding, Applications and Technology*, edited by M. Alexe, and U. Gösele, 85–106. Berlin, New York: Springer, 2004.
44. Agarwal, A., T. E. Haynes, V. C. Venezia, O. W. Holland, and D. J. Eaglesham. *Appl. Phys. Lett.* 72 (1998): 1086.
45. Nguyen, P., I. Cayrefourcq, K. K. Bourdelle, A. Boussagol, E. Guiot, N. Ben Mohamed, N. Sousbie, and T. Akatsu. *J. Appl. Phys.* 97 (2005): 083527.
46. Aspar, B., H. Moriceau, E. Jalaguier, C. Lagahe, A. Soubie, B. Biasse, A. M. Papon, et al. *J. Electron. Mater.* 30 (2001): 834.
47. Aspar, B., M. Bruel, H. Moriceau, C. Maleville, T. Poumeyrol, A. M. Papon, A. Claverie, G. Benassayag, A. J. Auberton-Hervé, and T. Barge. *Microelectron. Eng.* 36 (1997): 233.

48. Aspar, B., C. Lagahe, H. Moriceau, A. Soubie, M. Bruel, A. J. Auberton-Hervé, T. Barge, and C. Maleville. In *Symposium on Defects and Impurities in Semiconductors*, *Mater. Res. Soc. Symp. Proc.*, Vol. 510, 381. 1998.
49. Grisolia, J., G. Ben Assayag, A. Claverie, B. Aspar, C. Lagahe, and L. Laanab. *Appl. Phys. Lett.* 76 (2000): 852.
50. Akatsu, T., K. K. Bourdelle, C. Richtarch, B. Faure, and F. Letertre. *Appl. Phys. Lett.* 86 (2005): 181910.
51. Maleville, C., O. Rayssac, H. Moriceau, B. Biasse, L. Baroux, B. Aspar, and M. Bruel. *Semiconductor Wafer Bonding, Science Technology and Applications IV*, Electrochemical Society Proceedings, Vol. 97-36, 46–55. Pennington, NJ: Electrochemical Society, 1998.
52. Zheng, Y., S. S. Lau, T. Hochbauer, A. Misra, R. Verda, X.-M. He, M. Nastasi, and J. W. Mayer. *J. Appl. Phys.* 89 (2001): 2972–8.
53. Bourdelle, K. K., T. Akatsu, N. Sousbie, F. Letertre, D. Delpra, E. Neyre, N. Ben Mohamed, et al., *IEEE SOI Conference*, 98–9. October 2004.
54. Lagahe-Blanchard, C., N. Sousbie, S. Sartori, H. Moriceau, A. Soubie, B. Aspar, P. Nguyen, and B. Blondeau, In *Semiconductor Wafer Bonding VII: Science, Technology and Applications*, Proceedings Vol. 2003-19, 346. Pennington, NJ: The Electrochemical Society, 2003.
55. Nguyen, P., I. Cayrefourcq, B. Blondeau, N. Sousbie, C. Lagahe-Blanchard, S. Sartori, A.-M. Cartier. In *Proceedings of IEEE SOI Conference*, 132–34. 2003.
56. Uhlir, A. *Bell Syst. Technol. J.* 35 (1956): 333–47.
57. Baumgart, H., F. Phillipp, and G. K. Celler. In *Microscopy of Semiconducting Materials*, edited by G. Cullis, S. M. Davidson, and G. R. Booker, 223–8. London: Institute of Physics, 1983 Conference Series No. 67.
58. Yonehara, T., K. Sakaguchi, and N. Sato. *Appl. Phys. Lett.* 64 (1994): 2108.
59. Sakaguchi, K., and T. Yonehara. In *Wafer Bonding, Applications and Technology*, edited by M. Alexe, and U. Gösele, 107–56. Berlin, New York: Springer, 2004.
60. Izumi, K. *Vacuum* 42 (1991): 333.
61. Izumi, K., M. Doken, and H. Ariyoshi. *Electron. Lett.* 14 (1978): 593.
62. Hemment, P. L. F., K. J. Reeson, J. A. Kilner, R. J. Chater, C. Marsh, G. R. Booker, J. R. Davis, and G. K. Celler. *Nucl. Instr., Meth. Phys. Res.* B21 (1987): 129.
63. Jaussaud, C., J. Stoemenos, J. Margail, M. Dupuy, B. Blanchard, and M. Bruel. *Appl. Phys. Lett.* 46 (1985): 1064.
64. Celler, G. K., K. W. West, J. M. Gibson, and P. L. F. Hemment. *IEEE SOS/SOI Workshop*, Park City, Utah, 1985.
65. Celler, G. K., P. L. F. Hemment, K. W. West, and J. M. Gibson. *Appl. Phys. Lett.* 48 (1986): 532.
66. Celler, G. K., and A. E. White. *MRS Bull.* XVII, no. 6 (1992): 40.
67. Marsh, D., G. R. Booker, K. J. Reeson, P. L. F. Hemment, R. J. Chater, J. A. Kilner, J. A., Alderman, and G. K. Celler. In *Proceedings of European MRS Conference*, 137. 1986.
68. Krause, S., M. Anc, and P. Roitman. *MRS Bull.* 23, no. 12 (1998): 25.
69. Nakashima, S., and K. Izumi. *J. Mater. Res.* 8 (1993): 523.
70. Anc, M. J., R. P. Dolan, J. Jiao, T. Nakai. *IEEE International SOI Conference*, Rohnert Park, CA, 106. Piscataway, NJ: IEEE, 1999.
71. Nakashima, S., T. Katayama, Y. Miyamura, A. Matsuzaki, M. Kataoka, D. Ebi, M. Imai, K. Izumi, and N. Ohwada. *J. Electrochem. Soc.* 143 (1996): 244.
72. Holland, O. W., D. Fathy, and D. K. Sadana. *Appl. Phys. Lett.* 69 (1996): 674.
73. Davis, J. R., A. K. Robertson, K. J. Reeson, and P. L. F. Hemment. *Appl. Phys. Lett.* 51 (1987): 1419.
74. Hemment, P. L. F., A. K. Robertson, K. J. Reeson, J. R. Davis, J. A. Kilner, and J. Stoemenos. *Mater. Res. Soc. Symp. Proc.* 107 (1988): 87.
75. Cohen, G. M., and D. K. Sadana. *Mater. Res. Soc. Symp. Proc.* 686 (2002): A.2.4.1.
76. Ogura, A. In *Proceedings of IEEE International SOI Conference*, 185, Williamsburg, Piscataway, NJ: IEEE, 2002.
77. Dong, Y., M. Chen, J. Chen, X. Wang. In *Proceedings of IEEE SOI Conference*, 60–1. 2004.
78. Bean, K. E., and W. R. Runyan. *J. Electrochem. Soc.* 124 (1977): 5C.

79. Gupta, A., and P. K. Vasudev. *Solid State Technol.* 26 (1983): 104.
80. Pribat, D., L. M. Mercandalli, J. Siejka, and J. Perriere. *J. Appl. Phys.* 58 (1985): 313.
81. Leamy, H. J. *Mater. Res. Soc. Symp. Proc.* 4 (1982): 459.
82. Celler, G. K. *J. Cryst. Growth* 63 (1983): 429.
83. Fan, J. C. C., B.-Y. Tsaur, and M. W. Geis. *J. Cryst. Growth* 63 (1983): 453.
84. Celler, G. K., McD. Robinson, L. E. Trimble, and D. J. Lischner. *J. Electrochem. Soc.* 132 (1985): 211.
85. Jastrzebski, L. *J. Cryst. Growth* 63 (1983): 493.
86. Yamamoto, H., H. Ishiwara, and S. Furukawa. *Appl. Phys. Lett.* 46 (1985): 268.
87. Imai, K., and H. Unno. *IEEE Trans. Electron. Dev.* ED-31 (1984): 297.
88. Phillips, J. M. *Mater. Res. Soc. Symp. Proc.* 37 (1985): 143.
89. Jurczak, M., T. Skotnicki, M. Paoli, B. Tormen, J. Martins, J. L. Regolini, D. Dutartre, et al. *IEEE Trans. Electron Dev.* 47 (2000): 2179.
90. Bertrand, I., P. Renaud, J. M. Dilhac, and C. Ganibal. In *Proceedings of 7th International Seminar on Power Semiconductors (ISPS'04)*, Prague, August 31–September 3, 55–60. 2004.
91. Pretet, J., S. Monfray, S. Cristoloveanu, and T. Skotnicki. *IEEE Trans. Electron Dev.* 51, no. 2 (2004): 240–5.
92. Mazuré, C., and A.-J. Auberton-Hervé. In *Proceedings of ESSDERC 2005*, edited by G. Ghibaudo, T. Skotnicki, S. Cristoloveanu, and M. Brillouet, 29–38. Grenoble: IEEE, 2005.
93. Sato, T., Y. Takeishi, and H. Hara. *Phys. Rev.* B4 (1971): 1950.
94. Sayama, H., Y. Nishida, H. Oda, T. Oishi, S. Shimizu, T. Kunikiyo, K. Sonoda, Y. Inoue, and M. Inuishi. *IEDM* (1999): 657.
95. Matsumoto, T., S. Maeda, H. Dang, T. Uchida, K. Ota, Y. Hirano, H. Sayama, H. et al. *IEDM* (2002) paper 27-07.
96. Shang, H., J. Rubino, B. Doris, A. Topol, J. Sleight, J. Cai, L. Chang, et al. *VLSI*, 78–9. 2005.
97. Yang, M., M. Ieong, L. Shi, K. Chan, V. Chan, A. Chou, E. Gusev, et al. *IEDM*, 2003.
98. Doris, B., Y. Zhang, D. Fried, J. Beintner, O. Dokumaci, W. Natzle, H. Zhu, et al. In *Proceedings VLSI 2004 Conference*, 86–7, 2004.
99. Ieong, M., B. Doris, J. Kedzierski, Z. Ren, K. Rim, M. Yang, H. Shang, and L. Chang. *ECS Proc.* PV 2004-01 (2004): 371–82.
100. Langdo, T. A., A. Lochtefeld, M. T. Currie, R. Hammond, V. K. Yang, J. A. Carlin, C. J. Vineis, et al. In *Proceedings of 2002 IEEE International SOI Conference*, Williamsburg, VA, 211. Piscataway, NJ: IEEE, 2002.
101. Ghyselen, B., J.-M. Hartmann, T. Ernst, C. Aulnette, B. Osternaud, Y. Bogumilowicz, A. Abbadie., et al. *Solid State Electron.* 48 (2004): 1285–96.
102. Lee, M. L., E. A. Fitzgerald, M. T. Bulsara, M. T. Currie, and A. Lochtefeld. *J. Appl. Phys.* 97, no. 1 (2005): 011101.
103. Liu, C. W., S. Maikap, and C.-Y. Yu. *IEEE Circuits Dev.* 21, no. 3 (2005): 21–36.
104. Manasevit, H. M., I. S. Gergis, and A. B. Jones. *Appl. Phys. Lett.* 41 (1982): 464.
105. Dingle, R., H. L. Störmer, A. C. Gossard, and W. Wiegmann. *Appl. Phys. Lett.* 33 (1978): 665.
106. Green, M. L., B. E. Weir, D. Brasen, Y. F. Hsieh, G. Higashi, A. Feygenson, L. C. Feldman, and R. L. Headrick. *J. Appl. Phys.* 69 (1991): 745.
107. Welser, J., J. L. Hoyt, and J. F. Gibbons. *IEDM Tech. Dig.* (1992): 1000–2.
108. Welser, J., J. L. Hoyt, S. Takagi, and J. F. Gibbons. *IEDM Tech. Dig.* (1994): 373–6.
109. Fischetti, M. V., and S. E. Laux. *J. Appl. Phys.* 80 (1996): 2234–52.
110. People, R., and J. C. Bean. *Appl. Phys. Lett.* 47 (1985): 322.
111. Lyutovich, K., E. Kasper, and M. Oehme. *Mater. Res. Soc.* 809 (2004), paper B1.4.
112. Loo, R., R. Delhougne, P. Meunier-Beillard, M. Caymax, P. Verheyen, G. Eneman, I. De Wolf., et al. *Mater. Res. Soc.* 809 (2004), paper B1.2.
113. Mantl, S., R. Loo, R. Delhougne, P. Meunier-Beillard, M. Caymax, P. Verheyen, G. Eneman., et al. *Mater. Res. Soc.* 809 (2004), paper B1.6.
114. Tiberj, A., V. Paillard, C. Aulnette, N. Daval, K. K. Bourdelle, M. Moreau, M. Kennard, and I. Cayrefourcq. *Mater. Res. Soc. Proc.* (2004): 809.
115. Tezuka, T., N. Sugiyama, and S. Takagi. *Appl. Phys. Lett.* 79 (2001): 1798.

116. Nakaharai, S., T. Tezuka, N. Sugiyama, and S. Takagi. In *SiGe: Materials, Processing, and Devices*, Vol. 2004-07, edited by D. Harame, et al., *ECS Proceeding*, 741–8. Pennington, NJ: The Electrochemical Society, 2004.
117. Mizuno, T., N. Sugiyama, T. Tezuka, and S.-I. Takagi. *Appl. Phys. Lett.* 80 (2002): 601.
118. Zhiyuan, C., A. J. Pitera, M. L. Lee, J. Jongwan, J. L. Hoyt, D. A. Antoniadis, and E. A. Fitzgerald. *IEEE EDL* 25 (2004): 147–9.
119. Chui, C. O., H. Kim, D. Chi, B. B. Triplett, P. McIntyre, and K. Saraswat. *IEDM Tech. Dig.* (2002): 437–40.
120. Shang, H., H. Okorn-Schimidt, J. Ott, P. Kozlowski, S. Steen, E. C. Jones, H.-S. P. Wong, and W. Hanesch. *IEEE Electron. Dev. Lett* 24 (2003): 242–4.
121. F. Letertre, C. Deguet, C. Richtarch, B. Faure, J. M. Hartmann, F. Chieu, A. Beaumont, J. Dechamp, C. Morales, F. Allibert, P. Perreau, S. Pocas, S. Personnic, C. Lagahe-Blanchard, B. ghyselen, Y. M. Le Vaillant, E. Jalaguier, N. Kernevez, and C. Mazure. In *High-Mobility Group-IV Materials and Devices*, edited by M. Caymax, E. Kasper, S. Zaima, K. Rim, P.F.P. Fichter, (*Mater. Res. Soc. symp. Proc.* 809, Warrendale, PA, 2004), pp. 153–158.
122. Akatsu, T., C. Deguet, L. Sanchez, C. Richtarch, F. Allibert, F. Letertre, C. Mazure, et al. *IEEE SOI Conference*, Honolulu, Hawaii, October 2005.
123. Di Cioccio, L., E. Jalaguier, and F. Letertre. In *Wafer Bonding, Applications and Technology*, edited by M. Alexe, and U. Gösele, 263–314. Berlin, New York: Springer, 2004.
124. Maleville, C., and G. Celler. *Yield Manag. Solut. Mag.* Summer (2004): 6–11.
125. Maleville, C., W. McMilan, and A. Srivatsa. *Semicond. Int.* August (2004): 34.
126. Liu, R., and M. Canonico. *AIP Conf. Proc.* 683 (2003): 738.
127. Renucci, J. B., R. N. Tyte, and M. Cardona. *Phys. Rev.* B11 (1975): 3885–95.
128. Paillard, V., P. Puech, M. A. Laguna, P. Temple-Boyer, B. Caussat, J. P. Couderc, and B. de Mauduit. *Appl. Phys. Lett.* 73 (1998): 1718.
129. Moulin, C., D. Delprat, C. Maleville, W. McMillan, J. Payne, K. Birdwell, R. Brun, and R. Moirin. *IEEE SOI Conference*, Honolulu, Hawaii, October 2005.
130. Cristoloveanu, S., and S. Williams. *IEEE Electron. Dev. Lett.* 13, no. 2 (1992): 102–4.
131. Komiya, K., N. Bresson, S. Sato, S. Cristoloveanu, and Y. Omura. *IEEE Trans. Electron. Dev.* 52, no. 3 (2005): 406–12.
132. Cristoloveanu, S., D. Munteanu, and M. Liu. *IEEE Trans. Electron. Dev.* 47, no. 5 (2000): 1018–27.
133. Bresson, N., and S. Cristoloveanu. *Microelectron. Eng.* 72, no. 1–4 (2004): 357–61.
134. Jun, B., Y. V. White, R. D. Schrimpf, D. M. Fleetwood, F. Brunier, N. Bresson, S. Cristoloveanu, and N. H. Talk. *Appl. Phys. Lett.* 85, no. 15 (2004): 3095–7.
135. Hafez, I. M., G. Ghibaudo, and F. Balestra. *IEEE Trans. Electron. Dev.* 37 (1990): 818.
136. Jomaah, J., F. Balestra, and G. Ghibaudo. *Physica Status Solidi (a)* 142 (1994): 533.
137. Balestra, F., J. Jomaah, G. Ghibaudo, O. Faynot, A-J. Auberton-Hervé, and B. Giffart. *IEEE Trans. Electron. Dev.* 41 (1994): 109.
138. Yoshimi, M., M. Takahashi, T. Wada, K. Kato, S. Kambayashi, M. Kemmoshi, K. Natori., et al. *IEEE Trans. Electron. Dev.* 37 (1990): 2015.
139. Ver Ploeg, E. P., T. Watanabe, N. A. Kistler, J. C. S. Woo, and J. D. Plummer. *IEDM Tech. Dig.* (1992): 337.
140. Munteanu, D., D. A. Weiser, S. Cristoloveanu, O. Faynot, J.-L. Pelloie, and J. G. Fossum. *IEEE Trans. Electron. Dev.* 45, no. 8 (1998): 1678–83.
141. Lim, H. K., and J. G. Fossum. *IEEE Trans. Electron. Dev.* 30 (1983): 1244–51.
142. Omura, Y., T. Ishiyama, M. Shoji, and K. Izumi. In *SOI Technology and Devices VII*, 199. Pennington, NJ: Electrochemical Society, 1996.
143. Mazhari, B., S. Cristoloveanu, D. E. Ioannou, and A. L. Caviglia. *IEEE Trans. Electron. Dev.* 38 (1991): 1289.
144. Balestra, F., M. Benachir, J. Brini, and G. Ghibaudo. *IEEE Trans. Electron. Dev.* 37 (1990): 2303.
145. Ouisse, T., S. Cristoloveanu, and G. Borel. *Solid State Electron.* 35 (1992): 141.
146. Bawedin, M., S. Cristoloveanu, J. G. Yun, and D. Flandre. *Solid State Electron.,* 49, no. 9 (2005): 1547.

147. Balestra, F., S. Cristoloveanu, M. Bénachir, J. Brini, and T. Elewa. *IEEE Electron. Dev. Lett.* 8 (1987): 410.
148. Allibert, F., J. Pretet, G. Pananakakis, and S. Cristoloveanu. *Appl. Phys. Lett.* 84 (2004): 1192–4.
149. Zaouia, S., S. Goktepeli, A. H. Perera, and S. Cristoloveanu. In *Silicon-on-Insulator Technology and Devices XII*, edited by G. K. Celler, et al. *ECS Proceedings*, 309–16, Pennington, NJ: The Electrochemical Society, 2005.
150. Cristoloveanu, S., T. Ernst, D. Munteanu, and T. Ouisse. *Int. J. High Speed Electron. Syst.* 10, no. 1 (2000): 217–30.
151. Ernst, T., C. Tinella, and S. Cristoloveanu. *Solid State Electron.* 46, no. 3 (2002): 373–8.
152. Wong, H.-S., D. J. Franck, and P. M. Solomon. *IEDM Techn. Dig.* (1998): 407.
153. Yan, R.-H., A. Ourmazd, and K. F. Lee. *IEEE Trans. Electron. Dev.* 39, no. 7 (1992): 1704–10.
154. Franck, D., S. Laux, and M. Fischetti. *IEDM Techn. Dig.* (1992): 553.
155. Likharev, K. K. *Nano, Giga Challenges in Microelectronics*, 27–68. Amsterdam: Elsevier, 2003.
156. Doris, B., M. Ieong, H. Zhu, Y. Zhang, M. Steen, W. Natzle, S. Callegari, et al. *IEDM Technical Digest*, 27.3.1–4. Piscataway, NJ: IEEE, 2003.
157. Pretet, J., D. Ioannou, N. Subba, S. Cristoloveanu, W. Maszara, and C. Raynaud. *Solid State Electron.* 46, no. 11 (2002): 1699–707.
158. Elewa, T., B. Kleveland, S. Cristoloveanu, B. Boukriss, and A. Chovet. *IEEE Trans. Electron. Dev.* 39, no. 4 (1992): 874–82.
159. Majima, H., H. Ishikuro, and T. Hiramoto. *Technical Digest IEDM'99*, 379–82. 1999.
160. Pretet, J., A. Ohata, F. Dieudonn'e, et al. In *Silicon Nitride and Silicon Dioxide Thin Insulating Films VII*, Electrochemical Society Proceedings, Vol. PV-2003-02, 476–87. Pennington, NJ: Electrochemical Society, 2003.
161. Ohata, A., J. Pretet, S. Cristoloveanu, and A. Zaslavsky. *IEEE Trans. Electron. Dev.* 52, no. 1 (2005): 124–5.
162. Ernst, T., S. Cristoloveanu, G. Ghibaudo, T. Ouisse, S. Horiguchi, Y. Ono, Y. Takahashi, and K. Murase. *IEEE Trans. Electron. Dev.* 50 (2003): 830–8.
163. Fiegna, C., and A. Abramo, *International Conference on Simulation of Semiconductor Processes and Devices, SISPAD'97*, 93–6. 1997.
164. Gamiz, F., J. B. Roldan, A. Godoy, P. Cartujo-Cassinello, and F. Jimenez-Molinos. In *Silicon-On-Insulator Technology and Devices XI*, Electrochemical Society Proceedings, PV-2003-05, 343–8. Pennington, USA, 2003.
165. Ohata, A., M. Cassé, S. Cristoloveanu, and T. Poiroux. In *Proceedings of ESSDERC 2004*, 109–12. Piscataway, NJ: IEEE, 2004.
166. Chen, J., R. Solomon, T.-Y. Chan, P.-K. Ko, and C. Hu. *IEEE Electron. Dev. Lett.* 12 (1991): 453–5.
167. Chaisantikulwat, W., M. Mouis, G. Ghibaudo, C. Gallon, C. Fenouillet-Beranger, D. K. Maude, T. Skotnicki, and S. Cristoloveanu. In *Proceedings of ESSDERC 2005*, edited by G. Ghibaudo, T. Skotnicki, S. Cristoloveanu, and M. Brillouet, Grenoble: IEEE, 2005.
168. Pretet, J., T. Matsumoto, T. Poiroux, S. Cristoloveanu, R. Gwoziecki, C. Raynaud, A. Roveda, H. Brut. In *Procedings of ESSDERC'02*, 515–18. Bologna, Italy: University of Bologna, 2002.
169. Mercha, A., J. M. Rafi, E. Simoen, E. Augendre, and C. Claeys. *IEEE Trans. Electron. Dev.* 50, no. 7 (2003): 1675–82.
170. Dieudonné, F., S. Haendler, J. Jomaah, and F. Balestra. *Solid State Electron.* 48, no. 6 (2004): 985–97.
171. Cassé, M., J. Pretet, S. Cristoloveanu, T. Poiroux, C. Fenouillet-Beranger, F. Fruleux, C. Raynaud, and G. Reimbold. *Solid State Electron.* 48, no. 7 (2004): 1243–7.
172. Pretet, J., A. Vandooren, and S. Cristoloveanu. In *Proceedings of ESSDERC'03*, edited by J. Franca, and P. Freitas, 573–6. Estoril, Portugal: IEEE, 2003.
173. Su, L. T., K. E. Goodson, D. A. Antoniadis, M. I. Flik, and J. E. Chung. *IEDM Tech. Dig.* (1992): 357.
174. Bengtsson, S., M. Bergh, M. Choumas, C. Olesen, and K. O. Jeppson. *Jpn. J. Appl. Phys.* 35 (1996): 4175–81.
175. Oshima, K., S. Cristoloveanu, B. Guillaumot, H. Iwai, and S. Deleonibus. *Solid State Electron.* 48 (2004): 907–17.

176. Bresson, N., S. Cristoloveanu, C. Mazure, F. Letertre, H. Iwai. Solid State Electron., 49(9), 1522–8, 2005.
177. F. Gamiz, J.B. Roldan, J.A. Lopez-Villanueva, P. Cartujo-Cassinello, J. E. Carceller, and P. Cartujo. *Silicon-On-Insulator Technology and Devices X*, 157–68. Pennington: Electrochemical Society, 2001.
178. Esseni, D., M. Mastrapasqua, G. K. Celler, C. Fiegna, L. Selmi, and E. Sangiorgi. *Trans. Electron. Dev.* 50, no. 3 (2003): 802–8.
179. Cirba, C. R., S. Cristoloveanu, R. D. Schrimpf, L. C. Feldman, D. M. Fleetwood, and K. F. Galloway. In *Silicon-on-Insulator Technology and Devices XI*, Electrochemical Society Proceedings, Vol. 2003-05, 493–8. Pennington, USA: Electrochemical Society, 2003.
180. Hisamoto, D., W.-C. Lee, J. Kedzierski, E. Anderson, H. Takeuchi, A. Hideki, K. Asano, T-J. King, et al. *Technical Digest IEDM'98*, 1032–4. 1998.
181. Hisamoto, D., W.-C. Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, E. Anderson, et al. *IEEE Trans. Electron. Dev.* 47, no. 12 (2000): 2320–5.
182. Widiez, J., F. Daugé, M. Vinet, T. Poiroux, B. Previtali, M. Mouis, and S. Deleonibus. *IEEE International SOI Conference*, Charleston, USA, 2004.
183. Allibert, F., A. Zaslavsky, J. Pretet, and S. Cristoloveanu. In *Proceedings of SSDERC'2001*, 267–70 Frontier Group, 2001.
184. Chang, L., M. Ieong, and M. Yang. *IEEE Trans. Electron. Dev.* 51, no. 10 (2004): 1621–7.
185. Liu, Y. X., M. Masahara, K. Ishii, T. Sekigawa, H. Takashima, H. Yamauchi, and E. Suzuki. *IEEE Electron. Dev. Lett.* 25, no. 7 (2004): 510–2.
186. Xiong, W., G. Gebara, J. Zaman, M. Gostkowski, B. Nguyen, G. Smith, D. Lewis, et al. *IEEE Electron. Dev. Lett.* 25, no. 8 (2004): 541–3.
187. Daugé, F., J. Pretet, S. Cristoloveanu, A. Vandooren, L. Mathew, J. Jomaah, and B. -Y.Nguyen. *Solid State Electron.* 48 (2004): 535–42.
188. Cristoloveanu, S., S. R. Ritzenthaler, A. Ohata, and O. Faynot, *Intl J. High Speed Electron.* 16, no. 1 (2006): 9–30.
189. Skotnicki, T. In *Silicon-on-Insulator Technology and Devices X*, Vol. 2001-3, 391–402, Pennington: Electrochemical Society, 2001.
190. Fossum, J. G., J. W. Yang, and V. P. Trivedi. *IEEE Electron. Dev. Lett.* 24, no. 12 (2003): 745–7.
191. Blalock, B. J., S. Cristoloveanu, B. Dufrene, F. Allibert, and M. M. Mojarradi. "Frontiers in Electronics–Future Chips." *World Scientific* 26 (2002): 305–14.
192. Dufrene, B., K. Akarvardar, S. Cristoloveanu, B. J. Blalock, P. Gentil, E. Kolawa, and M. M. Mojarradi. *IEEE Trans. Electron Dev.* 51, no. 11 (2004): 1931–5.
193. Akarvardar, K., S. Cristoloveanu, P. Gentil, R. Schrimpf, and B. J. Blalock. *IEEE Trans. Electron Dev.* 54, no. 2 (2007): 323–31.
194. McConnell, A. D., S. Uma, and K. E. Goodson. *J. Microelectromech. Syst.* 10 (2001): 360–9.
195. Chu, P. B., S.-S. Lee, S. Park, M.-J. Tsai, I. Brenner, D. Peale, R. Doran, and C. Pu. *Proc. SPIE* 4561 (2001): 55–65.
196. Liu, A., R. Jones, L. Liao, D. Samara-Rubio, D. Rubin, O. Cohen, R. Nicolaescu, and M. Paniccia. *Nature* 427 (2004): 615–8.
197. Boyraz, O., and B. Jalali. *Opt. Exp.* 12 (2004): 5269–73.
198. Rong, H., R. Jones, A. Liu, O. Cohen, D. Hak, A. Fang, and M. Paniccia. *Nature* 433 (2005): 725–7.
199. Gunn, C. In *Proceedings of 2005 IEEE International SOI Conference*, 7–13. Piscataway, NJ: IEEE , 2005.
200. Kimerling, L. C. *Electrochem. Soc. Interface* (2000): 28–31.

5

Surface Preparation

5.1 RCA Clean **5**-2
5.2 Electrochemistry **5**-2
5.3 The Chemistry of Aqueous Solutions **5**-4
Solubility and Complexing Effect of pH • The Effect of Oxidation State
5.4 SC-1 ..... **5**-7
5.5 Complexing **5**-7
5.6 Particle Removal **5**-8
5.7 Particle Adhesion **5**-8
5.8 Particle Removal—Chemical Undercutting **5**-9
5.9 Particle Removal—Electrophoretic Effects and DLVO Theory **5**-10
5.10 Particle Removal—Megasonics **5**-12
5.11 Wet Chemical Etching **5**-15
5.12 Oxide Etch **5**-15
Dilute Hydrofluoric Acid Buffered Hydrofluoric Acid • Uses of Buffered HF
5.13 Hydrofluoric Acid and Metallic Contamination **5**-16
5.14 Defects Related to Drying **5**-17
5.15 Polysilicon Etch **5**-17
5.16 Selective Nitride Etch **5**-17
5.17 Oxide/Nitride Etch **5**-18
5.18 Bulk Organic Removal/Photoresist Strip **5**-18
5.19 Photolithography **5**-18
5.20 Sulfuric/Oxidizer Chemistry **5**-18
5.21 Reaction Mechanism **5**-19
5.22 Sulfuric/Peroxide (Piranha) **5**-20
5.23 DI/OZONE **5**-20
5.24 Surface Preparation and Cleaning for Interconnect **5**-21
5.25 Subtractive Aluminum Interconnect **5**-21
5.26 Key Subtractive Aluminum Cleaning Challenges with Associated Defects **5**-23
5.27 Copper/Low-*k* Dual Damascene Interconnect **5**-23
5.28 Key Cu/Low-*k* Challenges with Associated Defects **5**-23
5.29 Typical Defects **5**-28
5.30 Control and Monitoring of Surface Treatment Processes **5**-30
References **5**-32

Glenn W. Gale
SEZ AG

Brian K. Kirkpatrick
Texas Instruments, Inc.

Frederick W. Kern, Jr.
Hitachi Global Storage Technologies

Though the demise of wet cleaning has been predicted for many years, such cleans continue to be among the most prevalent operations in semiconductor processing. As devices and associated manufacturing flows have become increasingly complex, contaminants introduced by environment, machines, and process interactions have remained abundant, requiring wet cleans for their removal. New materials used to make devices have further introduced new and challenging contaminants. These trends are expected to continue.

Regarding the longevity of liquid-based cleaning steps, there are several reasons why these have not been replaced by dry processes. First, the affinity to dissolve contaminants is proportional to the number of solvent atoms or molecules that can be brought to bear. Thus a liquid is effective, since its density is on the order of a thousand times that of the corresponding gas. For example, a mole of liquid water occupies only 18 ml, compared with 22.4 l/mol of water vapor at standard temperature and pressure. So the solvents of choice are liquids. Additionally, liquids enable removal of metallic contaminants by complexation, and the fluid drag forces associated with liquids are highly useful in particle removal. Finally, electrophoretic effects in liquids can be controlled to aid in cleaning. These and other concepts will be discussed in detail in this chapter.

5.1 RCA Clean

The Radio Corporation of America (RCA) cleaning process is a mixture of aqueous ammonia and hydrogen peroxide (SC-1), followed by a mixture of hydrochloric acid and hydrogen peroxide (SC-2). First proposed by Dr Werner Kern [1] in the early seventies, it has been used as a general clean throughout the history of integrated circuit manufacturing. It continues to be used, with a reduction in temperature and concentration. This renders it less damaging to modern semiconductor device structures and optimizes the chemistry for single-pass operations rather than recirculated bath processing. Despite these changes, as well as the addition of megasonics to enhance particle removal and changes in some of the chemical species used, a base oxidizer mix followed by an acidified oxidizer are still the backbone of most cleaning operations.

5.2 Electrochemistry

Contaminants on the surface of wafers can be either inorganic or organic species. Organic hydrocarbons and elemental metals must be oxidized in order to be made soluble for removal from the wafer surface.

Redox (*Reduction/Oxidation*) reactions are the backbone of chemistry. Chemical reactions occur when electrons are transferred from the species being oxidized to the species being reduced. These reactions can be thought of in terms of an electrochemical cell, or battery. When dissimilar species are placed in an electrolyte and connected by an electrical conductor, a flow of electrons (current) is generated. The energy of this reaction can be measured as a cell voltage. The voltage created between any species and a reference reaction can be used to develop a table of oxidizing strengths. Moreover, if one considers the reactions taking place independently at the cathode (reduction reaction) and the anode (oxidation reaction), these reactions may be written as *half-cells*, with only the oxidation (or reduction) portion shown and the electron generated or consumed shown as a reactant, or reaction product, respectively. For example, the aqueous oxidation of Iron(II) is written:

$$Fe^0 + 2H^+ \rightarrow Fe^{+2} + H_2$$

This reaction can be separated into two half-reactions:

$$Fe^0 \rightarrow Fe^{+2} + 2e^- \quad \text{(oxidation)}$$

and:

$$2H^+ + 2e^- \rightarrow H_2 \quad \text{(reduction)}$$

As the above reaction (the reduction of the hydrogen ion to gaseous hydrogen) is the most prevalent reaction in aqueous chemistry driving the oxidation of species, it is arbitrarily assigned a half-cell voltage of zero, and all other half-cells are measured relative to this reaction. Doing this, results in a standard half-cell table. Standard half-cell potentials for species of interest for semiconductor processing are listed in Table 5.1.

By convention, all of these reactions are shown as *oxidation* half-cells. Obviously, in order to formulate a complete reaction, an oxidation *and a reduction* is required. A complete reaction is formed by reversing the direction of the half-cell for the species being oxidized (with the appropriate sign change for its cell voltage). To write a complete reaction, the number of electrons being transferred from the species oxidized to the species reduced must also be balanced stoichiometrically. As voltage is an intensive variable, it is not multiplied by the stoichiometric coefficient. Half-reactions, and their voltages, are added to give a complete reaction with a standard cell voltage. Those reactions with a net positive voltage proceed spontaneously, while those with a negative voltage do not occur or require the application of a counter-voltage (as in plating) to proceed. These are referred to as Standard Half-Cell Potentials, as they are valid for reactant and product concentrations of one mole per liter (or the partial pressure of a gas in equilibrium with a one mole per liter solution of any gaseous constituent) and 25°C temperature. Generally, the cell voltage increases linearly with the absolute temperature and varies logarithmically with the concentration of reactants (increases logarithmically) or reaction products (decreases logarithmically).

Useful information about reactions in aqueous solution can be gleaned from potential-pH diagrams, also known as Pourbaix diagrams [3,4]. These diagrams show the form of an element that will predominate, based on thermodynamic stability, at a particular solution pH and electrochemical potential. Water can act as either an oxidizing agent or a reducing agent. Low potentials indicate a reducing environment, while high potentials indicate an oxidizing environment. To remain stable in solution an element needs to avoid being oxidized or reduced. Overlaying the redox and acid-base chemistry of an element on the water stability diagram, results in the potential-pH diagram for that element. There are many examples of interest in wafer cleaning. Figure 5.1 shows the oxidizing strength of

TABLE 5.1 Standard Half-Cell Potentials for Species of Interest in Semiconductor Device Manufacturing

Reaction	E_0 (volts)
$Si^0 + 6OH^- \rightarrow SiO_3^{-2} + 3H_2O + 4e^-$	1.730
$Si^0 + 2H_2O \rightarrow SiF_6^{-2} + 4e^-$	1.200
$Si^0 + 2H_2O \rightarrow SiO_2 + 4H^+ + 4e^-$	0.840
$Zn^0 \rightarrow Zn^{+2} + 2e^-$	0.763
$Cr^0 \rightarrow Cr^{+2} + 2e^-$	0.557
$Fe^0 \rightarrow Fe^{+2} + 2e^-$	0.409
$Ni^0 \rightarrow Ni^{+2} + 2e^-$	0.230
$Fe^0 \rightarrow Fe^{+3} + 3e^-$	0.036
$Sn^0 \rightarrow Sn^{+2} + 2e^-$	0.136
$H_2 \rightarrow 2H^+ + 2e^-$	0.000
$Sn^{+2} \rightarrow Sn^{+4} + 2e^-$	−0.150
$Cu^0 \rightarrow Cu^{+2} + 2e^-$	−0.158
$4OH^- \rightarrow O_2 + 2H_2O + 4e^-$	−0.401
$Fe^{+2} \rightarrow Fe^{+3} + e^-$	−0.770
$NO_2 + H_2O \rightarrow NO_3^- + 2H^+ + e^-$	−0.940
$Cr^{+3} \rightarrow Cr^{+6} + 3e^-$	−1.100
$2H_2O \rightarrow O_2 + 4H^+ + 4e^-$	−1.229
$O_2 + OH^- \rightarrow O_3 + H_2O + 2e^-$	−1.240
$N_2 + 6H_2O \rightarrow 2NO_3^- + 12H^+ + 10e^-$	−1.250
$2H_2O \rightarrow H_2O_2 + 2H^+ + 2e^-$	−1.776
$O_2 + H_2O \rightarrow O_3 + 2H^+ + 2e^-$	−2.070

Source: Excepted from Weast, R. C., *CRC Handbook of Chemistry and Physics*, CRC Press, Boca Raton, FL, 1979, D155–D157.

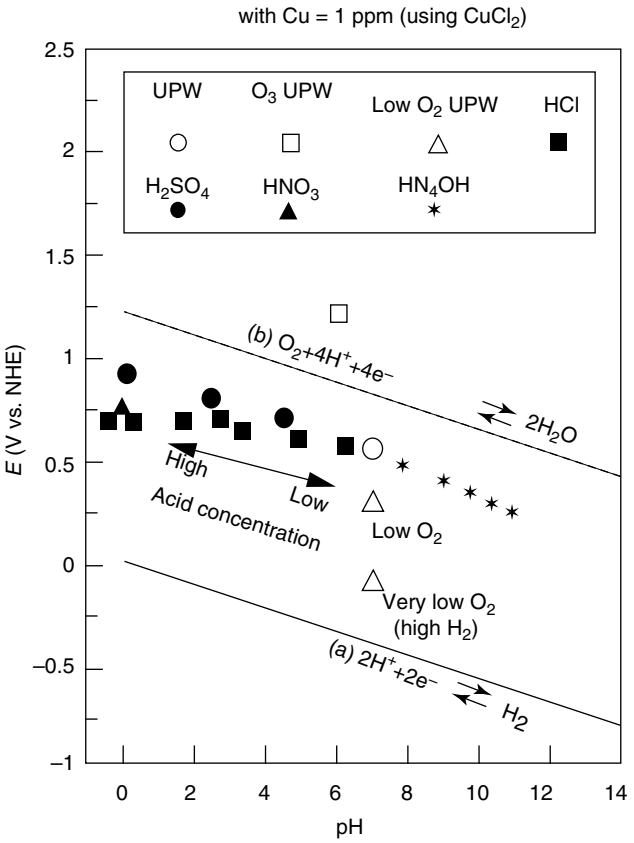


FIGURE 5.1 Potential-pH diagram for several common wafer cleaning chemicals. (From Ohmi, T., *ECS Proceedings*, PV95–20, The Electrochemical Society, Pennington, NJ, 1996, 1.)

various wafer cleaning chemicals in solution. It indicates, for example, that ozonated water has a stronger oxidation potential than even H_2SO_4 or HNO_3 [5]. The case of Cu, of interest since it is a harmful contaminant that is also abundant in semiconductor fabs where it is used for interconnect wiring, is discussed in detail in [6]. In the potential-pH diagram for Cu in water, we can see that Cu^{2+} (formed by oxidation of Cu) is favored at low pH, while cuprous and cupric oxide are favored near neutral pH, and $HCuO_2^-$ and CuO_2^- are formed by dissolution of copper oxides at alkaline pH levels [6]. Other examples showing or discussing Pourbaix diagrams in the semiconductor wafer cleaning literature include Hf and Zr [7], and As [8] (Figure 5.2).

5.3 The Chemistry of Aqueous Solutions

SC-1 is predominantly an organic clean. Hydrocarbons are oxidized by hydrogen peroxide to form carboxylic species (alcohols, aldehydes, and finally carboxylic acids). The presence of aqueous ammonia, which hydrolyzes to yield a basic solution when dissolved in water, helps to solvate these carboxylic species.

In the case of SC-2, hydrochloric acid both increases the oxidizing strength of the hydrogen peroxide by increasing the concentration of the reactant hydrogen ion (note the presence of the hydrogen ion as a reactant species in the above H_2O_2 reduction half-cell reaction) and serves as a complexing anion to increase the solubility of contaminant metals in the SC-2 solution.

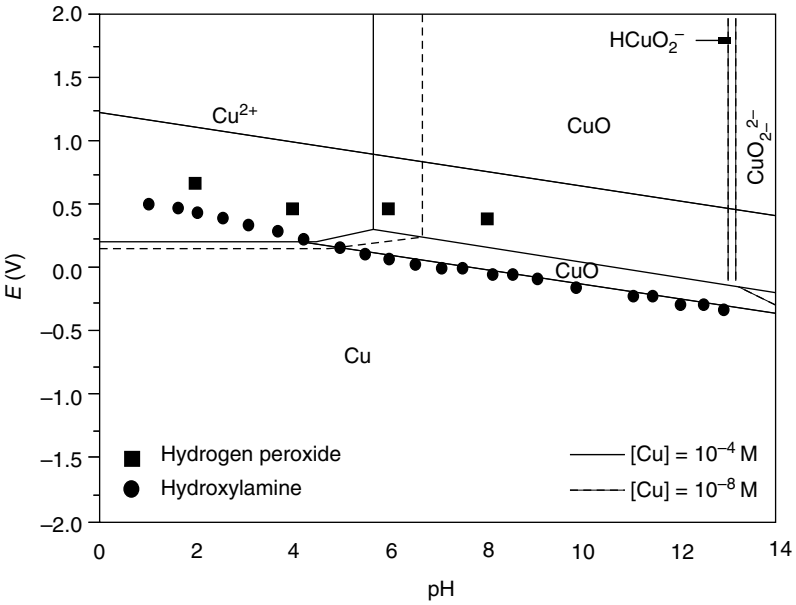


FIGURE 5.2 Potential-pH diagram for the Cu–H_2O system. (From Tamilmani, S., Huang, W., Raghavan, S., Small, R., *J. Electrochem. Soc.*, 149, 2002, G638–42.)

Inorganic species typically are solvated as ions. Ionic salts (NaCl, KCl, NH_4F, etc.) Are solvated from wafer surfaces directly, without the need for a redox reaction to take place. Metallic impurities are usually present on wafer surfaces as *elements* and therefore must undergo oxidation before metallic *ions* can be dissolved into solution. Non-noble metals (those with a positive half-cell potential—Fe, Cr, etc.) may be oxidized by the hydrogen ion present in water or the acidic SC-2 solution. Noble metals (e.g., Cu, Ag, Au) and Ni (which is usually found in the presence of non-noble metals e.g., with Fe and Cr in stainless steel) require the presence of a stronger oxidizer (H_2O_2) in order to be converted into an ionic state.

5.3.1 Solubility and Complexing

The amount of any ionic species that can be placed in solution is determined by its solubility expression. In general, for a species M_mX_x going into solution by ionizing into M^{+x} and X^{-m}, the solubility expression is:

$$K_{M_mX_x} = \frac{[M^{+x}]^m[X^{-m}]^x}{[M_m][X_x]}$$

Where $K_{M_mX_x}$ is a tabulated solubility constant, and $[M^{+x}]$, $[X^{-m}]$ and $[M_mX_x]$ are concentrations of the respective species in moles per liter. This assumes that M_mX_x is a soluble species (such as HF or acetic acid). If M_mX_x is a solid (e.g., NaCl, or $AlCl_3$), then the concentration of the solute is ignored, and the expression is what is known as the solubility product:

$$K_{sp} = [M^{+x}]^m[X^{-m}]^x$$

Typically with metals in solution, it is the formation of metal hydroxides that limits solubility. Table 5.2 lists K_{SP} values for a number of metal hydroxides commonly found and of concern in semiconductor processing.

TABLE 5.2 K_{sp} Values for Some Metal Hydroxides

Substance	Formula	K_{SP}
Aluminum hydroxide	$Al(OH)_3$	1.3×10^{-33}
Calcium hydroxide	$Ca(OH)_2$	5.5×10^{-6}
Chromium(II) hydroxide	$Cr(OH)_2$	2×10^{-16}
Chromium(III) hydroxide	$Cr(OH)_3$	6.3×10^{-31}
Copper(II) hydroxide	$Cu(OH)_2$	2.2×10^{-20}
Iron(II) hydroxide	$Fe(OH)_2$	8.0×10^{-16}
Iron(III) hydroxide	$Fe(OH)_3$	4×10^{-38}
Lead(II) hydroxide	$Pb(OH)_2$	1.2×10^{-15}
Lead(IV) hydroxide	$Pb(OH)_4$	3.2×10^{-66}
Magnesium hydroxide	$Mg(OH)_2$	1.8×10^{-11}
Nickel hydroxide	$Ni(OH)_2$	2.0×10^{-15}
Tin(II) hydroxide	$Sn(OH)_2$	1.4×10^{-28}
Tin(IV) hydroxide	$Sn(OH)_4$	1×10^{-56}
Titanium hydroxide	$Ti(OH)_3$	1×10^{-14}
Zinc hydroxide	$Zn(OH)_2$	6.3×10^{-49}

Source: From Dean, J. A., *Lange's Handbook of Chemistry*, McGraw Hill, New York, 1979, 5-7–12.

Clearly some metal hydroxides are very insoluble. While they would immediately precipitate if they were in solution, in actuality, they are not typically a problem, as their low solubility prevents their presence in the feedstock dionized (DI) water used to rinse wafers or formulate cleaning chemistries. Table 5.3 illustrates the solubility of some metal hydroxides as a function of pH and oxidation state.

5.3.2 Effect of pH

The solubility of all metal hydroxides is a strong function of pH. Most metal ions are multivalent. Their solubility is proportional to the *power* of the stoichiometric coefficient of the hydroxyl ion:

$$M(OH)_x \rightarrow M^{+x} + xOH^-$$

$$K_{SP} = [M^{+x}][OH^-]^x$$

TABLE 5.3 Metal Hydroxide Solubility as a Function of pH and Oxidation State

Substance	Formula	Solubility (ppm)		
		pH 5.3	pH 7.0	pH 9.74
Aluminum hydroxide	$Al(OH)_3$	3.72×10^{-3}	3.51×10^{-8}	7.83×10^{-21}
Calcium hydroxide	$Ca(OH)_2$	[1]	[1]	[1]
Chromium(II) hydroxide	$Cr(OH)_2$	NA	[1]	1.21×10^{-3}
Chromium(III) hydroxide	$Cr(OH)_3$	3.47	3.28×10^{-5}	3.80×10^{-18}
Copper(II) hydroxide	$Cu(OH)_2$	[1]	1.4×10^{6}	1.33×10^{-7}
Iron(II) hydroxide	$Fe(OH)_2$	NA	[1]	4.82×10^{-3}
Iron(III) hydroxide	$Fe(OH)_3$	[1]	[1]	2.41×10^{-25}
Lead(II) hydroxide	$Pb(OH)_2$	NA	[1]	7.23×10^{-3}
Lead(IV) hydroxide	$Pb(OH)_4$	7.02×10^{-35}	6.63×10^{-40}	4.18×10^{-50}
Magnesium hydroxide	$Mg(OH)_2$	[1]	[1]	[1]
Nickel hydroxide	$Ni(OH)_2$	[1]	[1]	0.0121
Tin(II) hydroxide	$Sn(OH)_2$	NA	0.02	8.44×10^{-16}
Tin(IV) hydroxide	$Sn(OH)_4$	1.26×10^{-25}	1.19×10^{-30}	6.03×10^{-44}
Titanium hydroxide	$Ti(OH)_3$	[1]	[1]	0.603
Zinc hydroxide	$Zn(OH)_2$	6.93×10^{-19}	6.54×10^{-24}	6.03×10^{-37}

NA, not applicable. Used for lower oxidation state species that would not exist in a strong oxidizer such as H_2O_2. [1], solubility dependent on associated anion, not hydroxide.

Source: From Dean, J. A., *Lange's Handbook of Chemistry*, McGraw Hill, New York, 1979, 5-49–53.

The last column of Table 5.3 (pH 9.74) depicts the effect of a basic solution on the solubility of metal hydroxides. Note, for example, Iron(III) Hydroxide, which goes from being essentially infinitely soluble in neutral water to a solubility of 2.41×10^{-25} ppm in a pH 9.74 medium. Conversely, acidifying a solution has the opposite effect. The solubility of Chromium(IV) Hydroxide increases by a factor of 100,000 by droping the pH of a solution from 7.0 to 5.33.

5.3.3 The Effect of Oxidation State

The effect of oxidation state on metal hydroxide solubilities is less evident than the effect of pH. Transition metals (Cr, Fe, Pb, Ni, and Sn in the previous table) can exist in multiple oxidation states. Without exception, the higher the oxidation state, the less soluble the metal hydroxide. For example, the solubility of Tin(IV) Hydroxide in neutral water is *38 orders of magnitude* lower than that of Tin(II) Hydroxide. In aqueous solution (without the presence of a strong oxidizer), most transition metals exist in a variety of their oxidation states. However, when a strong oxidizer (such as hydrogen peroxide, with a standard half-cell potential of 1776 V) is added, transition elements in a mix of oxidation states are each converted to their highest oxidation states.

5.4 SC-1

Increasing pH and oxidizing strength have a synergistic effect in the formulation of the SC-1 cleaning solution. The addition of basic aqueous ammonia to DI water and slightly acidic hydrogen peroxide can cause the precipitation of metal hydroxides from each of these solutions. Increasing the oxidizing strength (from the addition of hydrogen peroxide) can oxidize transition metals and lead to further metal hydroxide precipitation. The pH values of 5.3 and 9.74 were not chosen arbitrarily in the preceding table, rather they are the pH values corresponding to a 30% hydrogen peroxide solution and the SC-1 mix [10], respectively. The metals highlighted in this table (Cr, Cu, Fe, Pb, Ni, and Sn) can all lead to metal hydroxide precipitation when SC-1 is mixed, as demonstrated by their varying solubility limits. For example, Chromium(III) Hydroxide has a solubility of 3.47 ppm in hydrogen peroxide, and Chromium(II) Hydroxide is soluble in DI water. When mixed with hydrogen peroxide, the soluble Chromium(II) is oxidized to Chromium(III), dropping the solubility to 3.47 ppm. When mixed with aqueous ammonia, the solubility drops further to 2.07×10^{-15} ppm. All of the chromium that was stable in the DI and hydrogen peroxide precipitates out as $Cr(OH)_3$.

5.5 Complexing

Hydrochloric acid is used in SC-2 to both increase the oxidizing strength, and to complex metals. Complexing is a common occurrence with most transition metals. Neutral species (such as aqueous NH_3) or anions (Cl^-, F^-, etc.) can form coordination bonds with the numerous valence electrons of metal cations. The result is a new ion that effectively removes the amount of the complexed species from the equilibrium expression—thus, allowing much higher solubility levels to be reached. For example, Iron(III) can form the following complexes with the chloride ion:

$$FeCl^{+2}$$

$$FeCl_2^+$$

$$FeCl_3$$

$$FeCl_4^-$$

From the previous discussion, at a pH of 2.63 (1:1:20 $HCl:H_2O_2:H_2O$), $Fe(OH)_3$ has a solubility of only 5.15×10^{-4} mol/l (29 ppm). Taking the formation of $FeCl^{+2}$, $FeCl_2^+$ and $FeCl_4^-$ into account, in

TABLE 5.4 Complexing of Metals by Species Typically Used in Semiconductor Cleaning Operations

Aqueous Ammonia (NH_3)	Chloride (Cl^-)	Fluoride (F^-)	Hydroxide (OH^-)
Cobalt(I)	Cadmium	Aluminum	Aluminum
Cobalt(II)	Copper(II)	Chromium(III)	Cadmium
Copper(II)	Iron(II)	Iron(III)	Iron(II)
Iron(II)	Iron(III)	Magnesium	Iron(III)
Nickel	Lead		Lead(II)
Zinc	Silver(I)		Nickel
	Tin(II)		Zinc
	Tin(IV)		
	Zinc		

Source: From Dean, J. A., *Lange's Handbook of Chemistry*, McGraw Hill, New York, 1979, 5-49–53.

addition to the solubility of Fe^{+3} (as $Fe(OH)_3$), the capacity of Fe(III) spread among all of the soluble species is about 0.3 M (almost three orders of magnitude higher, and limited by the chloride present).

Tables of complexes can be found in a number of chemistry reference texts (e.g., Table 5.5 through Table 5.14 in Lange's Handbook of Chemistry). These typically list p*K* values for the various species formed, allowing one to compute solubility limits for metal species. A list of metals complexed by species typically used in semiconductor processing is shown in Table 5.4.

5.6 Particle Removal

Particulate contamination in semiconductor processing is a major cause of yield loss. Contaminant particles can locally mask lithographic, implant, or etch steps, cause shorts and opens, and degrade oxide integrity. Particles with characteristic dimensions just a fraction of the chip's smallest feature size can lead to killer defects [12]. The critical particle size for a technology has been considered by the International Technology Roadmap for Semiconductors (ITRS) to be 1/2 the first metal level half pitch [13]. Thus, the number of particles capable of causing killer defects is equal to the *cumulative* population equal to or larger than the critical particle size.

When surface particle density at a desired particle size is not known but data at other sizes are available, a useful extrapolation based on a *one over x-squared* relationship can be applied. Historically, this comes from the representation of an airborne particle size distribution where the number of particles of any one size (volume, or l^3) is proportional to the inverse of that particle size. Hence, the cumulative particle size distribution (the integral of $1/l^3$) is proportional to $1/x^2$, where x represents the critical particle size.

Particle removal in front-end processing is accomplished principally using SC-1 solutions, often assisted by the addition of sonic energy to the solution (megasonics) or, more recently, specialized jet sprays. The mechanisms contributing to particle removal are chemical undercutting of particles, electrophoretic effects, and physical effects. These will be described after a brief explanation of particle adhesion in liquid media.

5.7 Particle Adhesion

In a liquid, the adhesion of a particle to a substrate is typically dominated by the Van der Waals force [14], the sum of individual molecular forces given by [15]:

$$F_{\text{vdw}} = \frac{AR}{6H_0^2}$$

where R is the particle radius, H_0 is the separation distance between the particle and the wafer (typically taken to be 0.4 nm), and A is called the Hamaker constant. The Hamaker constant is a function of the two interacting materials and the medium separating them.

The Van der Waals force must be overcome in order for a particle to be removed from a wafer surface. It is worth noting that this force decreases linearly with decreasing particle size, whereas many physical removal mechanisms, such as shear stress, have forces proportional to the radius squared. Thus, particle removal becomes more difficult as size decreases. Additionally, capillary forces can in some cases contribute to particle adhesion. The capillary force, F_{cap}, is given by [5,16]:

$$F_{cap} = 4\pi R\gamma$$

where γ is the surface tension of the liquid.

Finally, particles can sometimes be chemically bonded to wafer surfaces (an example is chemical–mechanical polishing slurry). Such particles can be considerably more difficult to remove.

5.8 Particle Removal—Chemical Undercutting

Particles adhered to a silicon wafer can be undercut by slight etching of the surface beneath them. An SC-1 solution will simultaneously oxidize Si to form SiO_2 and etch SiO_2, resulting in gradual consumption of Si. The etching of SiO_2 in SC-1 follows [17]:

$$SiO_2 + 4OH^- \rightarrow SiO_4^{-4} + 2H_2O \quad \text{(soluble silicates)}$$

Silicon is also oxidized, then etched by the hydroxide ion [17]:

$$Si + 2OH^- + H_2O \rightarrow SiO_3^{-2} + 2H_2$$

Historically, the SC-1 solution has comprised a 5:1:1 by volume mixture of H_2O:30% H_2O_2:29% $NH_3 \cdot H_2O$. In such solutions, Si will not be directly etched because it is oxidized by H_2O_2 [18]. If the solution composition is not well controlled, however, and the $NH_3 \cdot H_2O$ volume ratio is allowed to significantly exceed that of H_2O_2, the SiO_2 etch rate can increase and direct Si attack can occur. During the lifetime of a process bath, there will be evaporation of ammonia and decomposition of hydrogen peroxide according to [19]:

$$2H_2O_2 \rightarrow 2H_2O + O_2$$

If the H_2O_2 decomposition rate is too fast, the etching behavior of the solution will no longer be controllable. Metal contaminants such as iron in the form of $Fe(OH)_3$ in the solution will catalyze H_2O_2 decomposition [18]. Today's state-of-the-art wet benches feature chemical makeup or "spiking" of $NH_3 \cdot H_2O$ and H_2O_2, the rate of which can be programmed based on experimental determination during tool setup or based on real-time monitoring of H_2O_2 and $NH_3 \cdot H_2O$ concentrations available with some equipment. However, the formation of water through the disproportionation of H_2O_2 (above) leads to dilution of the reactants and limits the number of times that spiking can return a bath to the desired reactant concentrations.

Another deleterious effect of H_2O_2 decomposition is its acceleration of Si surface roughening when SC-1 is used following an HF step. The degree of roughening has been directly correlated with the rate of evolution of O_2 from decomposition of H_2O_2 [19]. One hypothesis is that O_2 bubbles sticking to hydrophobic Si locally micromask the surface, which leads to non-uniform etching and, therefore, roughness. Figure 5.3 shows the increase in roughening, measured as laser light point defects (LPDs), as a function of O_2 evolution rate in SC-1. Light point defects in large size ranges represent particles; whereas,

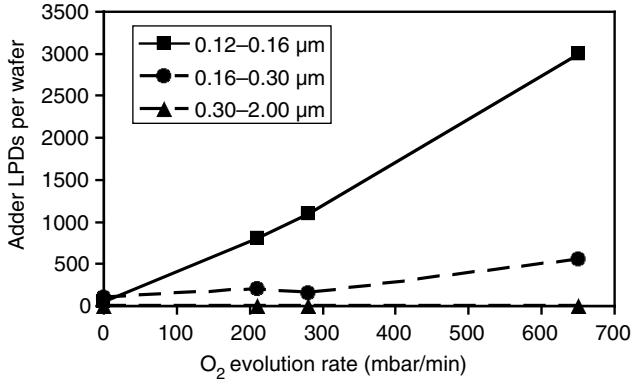


FIGURE 5.3 Surface roughening (measured as light point defects (LPDs) in the size range of 0.12–0.16 μm) as a function of O_2 evolution rate from H_2O_2 decomposition in SC-1. (From Schmidt, H. F., Meuris, M., Mertens, P. W., Rotondaro, A. L. P., Heyns, M. M., Hurd, T. Q., Hatcher, Z., *Jpn. J. Appl. Phys.*, 34, 727, 1995.)

LPDs in the range of 0.12–0.16 μm in the figure represent surface roughening [19]. Metals such as Fe and Cu, since they catalyze peroxide decomposition, also lead to increased roughening.

Schmidt has also determined that there is an incubation period that is required before bubbles are evolved as a result of metal-catalyzed decomposition. This time is in excess of that normally required to carry out a single SC-1 cleaning step [18]. For this reason, in addition to purity and the trend toward more diluted chemicals, single-pass cleaning (such as in spray or displacement processing) has become a technique preferred to traditional recirculated baths. Furthermore, the use of chemical mixes significantly reduced from the customary 5:1:1 ratio, has contributed to both economic and environmental feasibility of these approaches.

Temperature, since its increase generally results in increased chemical reaction rates, also significantly affects the rate of H_2O_2 decomposition and, thus, roughening [20]. In addition, the spiking of H_2O_2 during an SC-1 bath life can also result in increased roughening [19]. Therefore, the industry has shifted from traditional 5:1:1 SC-1 chemistries at 70°C–80°C to more dilute mixtures (10:1:1 down to about 50:1:1) at lower temperatures. It has been shown that these less aggressive chemistries can be used with no ill effects on particle removal efficiency (PRE), particularly when assisted by megasonic energy.

5.9 Particle Removal—Electrophoretic Effects and DLVO Theory

Solid materials immersed in electrolytic media acquire a charge, typically through adsorption of ions from solution (for example, OH^-), or by dissociation or ionization of surface groups. A layer of counter-ions in the liquid, charged oppositely to the surface charge, is attracted to the surface. This layer of charged species, called the Stern layer, moves with the surface (for example, a solid particle moving in a liquid) and repels ions of the same charge beyond it. Thus, a diffuse layer of ions form in the surrounding liquid. The two layers are collectively termed the electrical double layer, and the boundary between them is the shear plane [21]. Thus, in the case of a particle, the Stern layer moves with the particle, while the diffuse layer moves with the bulk of the liquid. An electric potential gradient exists between the particle surface and the bulk liquid; the potential at the shear plane is called the zeta potential. The movement of particles in a fluid due to charge effects is termed *electrophoresis*.

Although colloid science has applied these principles to the stabilization of fine particles in liquid media for many years, it took considerable time to recognize that the same phenomena occur at all liquid/solid interfaces and to apply these principles to the decontamination of semiconductor wafer surfaces [22].

The zeta potential largely determines the nature of electrophoretic motion of a particle in a liquid and how it will interact with other surfaces in the liquid. A particle's zeta potential is a strong function of both the pH and the ionic strength of the solution. It can be determined from electrophoretic mobility measurements, in which the velocities of particles in an electrolyte with an applied potential gradient are measured, typically using laser doppler velocimetry [23].

If the zeta potentials of a particle and a wafer surface are of opposite signs, the particle will be attracted to the wafer surface. If their zeta potentials are of the same sign, the particle will be repelled from the wafer surface. The zeta potential of a large surface such as a silicon wafer is most appropriately measured using the streaming potential technique [24].

The increase in OH^- ion concentration with increasing pH leads to a corresponding decrease in zeta potential from a positive value at low pH to a negative value at high pH. The pH at which the zeta potential is zero is termed the isoelectronic point, or point of zero charge [15,16]. For example, the zeta potentials of silica (SiO_2) and silicon nitride (Si_3N_4) particles, common contaminants in semiconductor process baths, are shown as a function of pH in Figure 5.2. One can see that the zeta potential reaches large negative values at high pH (resulting from the release of H^+ ions into solution). This indicates that particles will tend to be repelled from the negatively charged oxidized (hydrophilic) silicon wafer surface in highly basic solutions. This electrical repulsion of particles from wafers at high pH contributes to the efficacy of SC-1 solutions (which are typically of pH 9–10 depending on their concentration) for particle removal.

The ionic strength of the solution affects the distances over which electrophoretic interactions take place. The Debye length, κ^{-1}, is considered to be the approximate thickness of the electrical double layer. For aqueous solutions at 25°C, its reciprocal can be taken as [25]:

$$\kappa^{-1} = 2.32 \times 10^9 (2I)^{1/2}$$

where I is the solution ionic strength in moles/liter, and κ is in m^{-1}. This equation indicates that, as ionic strength increases, the debye length decreases and particles need to be closer together in order for their electrical double layers to interact. Consequently, double layer repulsion can be somewhat enhanced in more dilute solutions.

The different behavior of SiO_2 and Si_3N_4 particles illustrated in Figure 5.4 is also noteworthy. Silica particles have an isoelectric point of about pH 3 and a large negative zeta potential at high pH; whereas, Si_3N_4 particles have an isoelectric point of about pH 8 and a less negative zeta potential at high pH. Thus, in neutral pH DI water, silica particles are repelled from a hydrophilic wafer surface; whereas, silicon nitride particles are slightly attracted. In SC-1 solutions, the repulsion between particle and wafer is stronger for SiO_2 particles than for Si_3N_4. Silicon nitride particles, therefore, represent an effective challenge for cleaning processes and are sometimes used for testing purposes [26]. However, any such

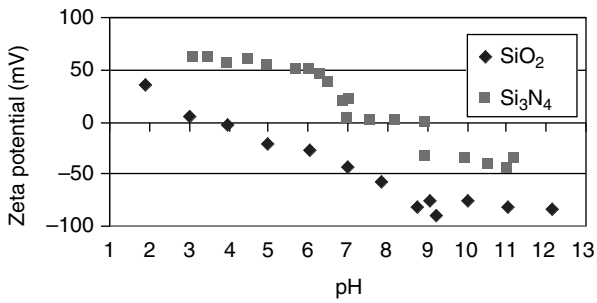


FIGURE 5.4 Zeta potential of SiO_2 and Si_3N_4 particles at 10^{-3} M ionic strength as a function of pH. (From Schmidt, H. F., Meuris, M., Mertens, P. W., Rotondaro, A. L. P., Heyns, M. M., Hurd, T. Q., Hatcher, Z., *Jpn. J. Appl. Phys.*, 34, 727, 1995.)

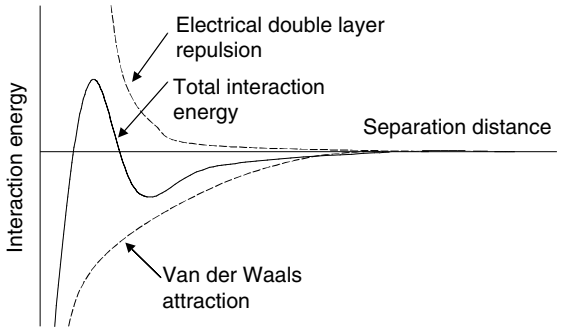


FIGURE 5.5 Energy of attraction and repulsion as a function of separation distance between surfaces, in accordance with DLVO theory.

experiment is quite sensitive to other factors such as the particle deposition method and the storage time of samples between deposition and cleaning.

The overall energy of interaction between two surfaces in a liquid medium, combining electrical double layer and Van der Waals effects, is described by DLVO theory (from the researchers Derjaguin B., L. D. Landau, E. J. W. Verwey, and J. Th. G. Overbeek) [14]. At small separation distances, the Van der Waals force dominates, while double layer repulsion governs the interaction at large separations. The total interaction is as shown in Figure 5.5. There is an energy barrier, due to double layer repulsion, which must be overcome in order for adhesion to occur. If the double layer interaction is attractive, there is no such barrier to adhesion.

5.10 Particle Removal—Megasonics

Megasonic cleaning [23,27–32] is an extension of ultrasonic cleaning, familiar to many for its use in removing contaminants from jewelry, tools, or other small parts. The term ultrasonic merely refers to sound waves transmitted at a frequency above the range of human hearing (greater than about 17 kHz). Ultrasonic cleaning baths typically operate at about 20–100 kHz. A sound wave is a pressure wave, causing pressure to vary sinusoidally with position and time according to:

$$p(x,t) = p_0 \sin(kx - \omega t + \phi)$$

where x is position, t is time, p_o is the pressure amplitude, k is the angular wave number (equal to 2π divided by the wavelength), ω is the angular frequency (equal to 2π times the frequency in Hz), and ϕ is a phase shift. When ultrasound waves are transmitted through a liquid, the alternation of compression and rarefaction leads to a "pulling apart" of the liquid during the low pressure phase to form cavities. Gas dissolved in the liquid diffuses rapidly into cavities, and they grow over brief time scales before collapsing during the high pressure phase of sonication. This phenomenon, known as cavitation, initiates a range of effects. For example, the oscillation of cavities (typically gas-filled) around an equilibrium radius, termed stable cavitation, creates flows which are highly localized but also associated with large velocity gradients. This so-called microstreaming can contribute to particle removal. A more severe form of cavitation, transient cavitation, is associated with violent implosion of mainly vapor-filled cavities and has been shown [28] to be capable of damaging surfaces. In recent years pattern damage, particularly to narrow polysilicon gate structures, has presented a major challenge for wafer cleaning engineers.

The low frequencies associated with ultrasonic cleaning would have been too damaging for semiconductor wafer cleans even many technology generations ago. However, when RCA researchers

[27,33] used sonic energy in baths at higher frequencies, near 1 MHz, they found that removal of particles could be accomplished without damage to semiconductor wafers. Their invention was called megasonic cleaning. Cavitation initiation is a strong function of frequency. As frequency is increased, the pressure amplitude (proportional to power input to the transducers) required for both stable and transient cavitation to occur, increases significantly. Experiments have determined [23,34–37] that cavitation does in fact occur in megasonic cleaning. How to harness the particle removal ability of cavitation while avoiding cavitation events that can damage patterns, is a topic of several current investigations.

In addition to cavitation, flow patterns resulting from attenuation of wave energy due to viscosity are generated in sonic baths. This phenomenon, called acoustic streaming, can aid in rapidly sweeping away particles already detached from surfaces. These particles flow out of the tank and are filtered in a typical recirculating tank system or disposed down the drain if the bath is nonrecirculating. In a typical overflow bath, the modest flow velocity produces a thick hydrodynamic boundary layer in which velocity between wafers in a process cassette varies parabolically from zero at the wafer surface to about 1.5 cm/s at the midpoint between wafers according to:

$$v(x) = 4V_{\max}(x/h)(1 - x/h)$$

where v is the velocity at any point x distance from a wafer surface, $V_{\max}$ is the velocity at the midpoint between wafers, and h is the separation distance between the wafers.

Thus, the regions close to the surface ($x \rightarrow 0$) have very small velocities, and transport is dominated by diffusion. Diffusivity of a contaminant is given by the Stokes–Einstein equation [14]:

$$D = kT/6\pi\mu R$$

where k is Boltzmann's constant, μ is the viscosity and T is absolute temperature. For 0.5 μm particles in water, this leads to diffusivities on the order of only 10^{-8} cm^2/s.

In a sonicated bath, however, acoustic streaming enhances transport. The boundary layer associated with acoustic streaming has a thickness of:

$$d_{\text{ac}} = (2\nu/\omega)^{1/2}$$

where ν is the kinematic viscosity. For example, at 850 kHz the acoustic boundary layer is only about 0.6-μm thick.

It has been found that the use of megasonics facilitates particle removal with considerably less need for undercutting. For several years, the minimum SiO_2 surface etching requirement for highly efficient particle removal with megasonics using methods available was about 3–12 Å [38]. In recent years, however, oxide, and Si loss from cleaning steps after forming the transistor gate stack has become a critical issue in front end of line (FEOL) cleaning. In particular, with increasingly shallow junctions, Si loss in the source-drain extension regions–accumulating over several cleaning steps–leads to decreased drive current [13]. Accordingly, process optimization has reduced the amount of Si and oxide loss in some wet bench megasonic SC-1 processes [39]. Use of more dilute, lower temperature SC-1 solutions leads to less Si surface roughening and less etching of oxide films. Particle removal in dilute solutions with megasonics has been shown to be very efficient, which may be partly due to the higher electrical double layer thickness at lower ionic strengths as described above. However, because the actual targets as defined by the ITRS require high PRE with <0.5 Å Si and SiO_2 loss per cleaning step, with no pattern damage, uniformity is critical and increased attention is being given to single wafer solutions. Fundamentally it is easier to control single wafer systems, where the uniformity problem can be largely reduced from three to two dimensions, compared with batch tanks in which different wafers and different regions of a wafer see different conditions in a complex megasonic field. This has enabled the use of more dilute, lower temperature SC-1 solutions, which lead to less Si surface roughening and less etching of oxide films. Particle removal in these dilute solutions with megasonics has been shown to be

TABLE 5.5 Effect of SC-1 Dilution on pH, Ionic Strength, and Hydroxide Ion Concentration

SC-1 Ratio	pH	I (mM)	$[OH^-]$ (mM)
5:1:1	9.74	72.3	0.815
5:1:0.5	9.59	54.4	0.548
5:0.5:0.5	9.75	41.5	0.758
5:1:0.3	9.48	43	0.408
5:0.3:0.3	9.75	26.6	0.715
5:1:0.1	9.23	25.6	0.214
5:0.1:0.1	9.74	9.63	0.631

Source: Data determined theoretically based on data from Rath, D. L., Unpublished SC-1 modeling work.

very efficient, which may be partly due to the higher electrical double layer thickness at lower ionic strengths as described above.

Table 5.5 illustrates the effects of SC-1 dilution on solution pH, ionic strength, and OH^- ion concentration. It is clear from the table that, in this particular dilution range, ionic strength decreases when the solution is diluted; whereas, pH remains nearly constant as long as a constant 1:1 ratio of H_2O_2:$NH_3 \cdot H_2O$ is maintained. The table also illustrates how $[OH^-]$, which determines the etch rates of SiO_2 and Si, depends more on $[NH_3 \cdot H_2O]$ and less on $[H_2O_2]$.

Removal of particles from silicon wafers in SC-1 megasonic baths can be summarized as follows. A combination of cavitation (physical) and slight surface etching (chemical) effects serve to detach particles from wafer surfaces. The high solution pH repels detached particles from the wafers and prevents redeposition. Acoustic streaming aids in transporting the detached particles away from the wafers toward a filter or drain.

Particle removal can also be accomplished by other, less common, means. Researchers at the Interuniversity Microelectronics Center (IMEC) in Belgium, who have been largely responsible for elucidating the science and mechanisms of the SC-1 clean, used similar principles to develop the IMEC Clean [40], a sequence which utilizes a sulfuric acid-ozone mixture (SOM), or a simpler ozonated water treatment, followed by a dilute HF and HCl treatment. Instead of simultaneously oxidizing and etching to undercut particles as in SC-1, the IMEC Clean first grows a chemical oxide in the ozonated bath, then removes it in the HF bath. A clean chemical oxide can then be optionally grown to re-passivate the surface. The technique has been demonstrated to result in very little silicon consumption, surface roughness, or metallic contamination [40].

Another chemical technique, known as single wafer cleaning with repetitive use of ozonized water (SCROD) and dilute HF (DHF), has been developed by sony researchers. Very short steps of O_3/H_2O (oxidizer) and 1% dilute HF (etchant) are alternated, using a single wafer spin processor, to remove particles by undercut and centrifugal force [41]. Since megasonic energy is not applied, pattern damage is not a concern. However, consumption of the substrate is required to effect particle removal by this method.

Other physical techniques have also been developed as alternatives to megasonics, due mainly to pattern damage concerns. A two-fluid jet spray [42] using nitrogen and DI water, creates a mist of small liquid droplets that are accelerated to the wafer surface. It is hypothesized that rapid spreading of the droplets upon impact effects particle removal.

A dry technique for particle removal uses a high velocity spray of cryogenic aerosol particles to dislodge contaminant particles by momentum transfer without harm to underlying structures on the wafer [43].

For particle removal in the back-end of the line, where metal wiring is present, SC-1 solutions cannot be used because they will attack the metal. Back end of line cleaning is discussed in detail elsewhere in this chapter.

TABLE 5.6 Common Etchants Used in Semiconductor Manufacturing

Film	Etchant
Silicon oxides (SiO_2)	Dilute HF (DHF)
	Buffered HF (BHF)
Polysilicon	Alkaline hydroxide + organic
Silicon nitride (Si_3N_4)–selective to SiO_2	Boiling phosphoric acid (H_3PO_4)
Silicon nitride/Silicon oxide (non-selective)	Hydrofluoric acid + organic

5.11 Wet Chemical Etching

Another traditional application of wet processing in semiconductor manufacturing has been etching of thin films. Table 5.6 shows etchants which have commonly been used to remove typical films.

5.12 Oxide Etch

The etching of silicon oxides by hydrofluoric acid solutions are among the most prevalent and time-honored reactions in semiconductor processing. Silicon dioxide (SiO_2) is rendered soluble by conversion to the silicon hexafluoride ion ($SiF_6^{=}$) by hydrofluoric acid. The traditional net reaction for this is depicted below:

$$SiO_2 + 4H^+ + 6F^- \tau SiF_6^{=} + 2H_2O$$

However, many of the properties of hydrofluoric acid cannot be accounted for by using the model of a simple binary, haloacid (HF). This would predict that, like HCl, HF should be a fully-dissociated strong acid in aqueous solution. In fact, its pK value for the dissociation constant is only 3.18 ($K_a = 6.61 \times 10^{-4}$) [44], demonstrating that HF is a weak acid. This has been attributed to hydrogen bonds forming to create aggregates of the form $(HF)_x$. The mechanism of HF dissociation in water has been shown by spectrographic studies [45] to be:

$$HF + H_2O \rightarrow [H_3O^+F^-] \rightarrow H_3O^+ + F^-$$
$$HF + [H_3O^+F^-] \rightarrow H_3O^+ + HF_2^-$$

5.12.1 Dilute Hydrofluoric Acid

Hydrofluoric acid undergoes dissociation to form *equal amounts* of hydrogen and fluoride ions. As DHF reacts with SiO_2, *more* fluoride than hydrogen is consumed. Thus, as this reaction proceeds, the etching solution becomes more acidic, which in turn suppresses the fluoride ion concentration, owing to the solubility relationship for hydrofluoric acid:

$$K = \frac{K_a[HF]}{[H^+]}$$

5.12.2 Buffered Hydrofluoric Acid

Buffered HF (BHF) is a mixture of ammonium fluoride (NH_4F) and hydrofluoric acid. As with any combination of an acid and its salt of a weak base, the resulting mixture is what is known as a buffer solution; hence, the name Buffered HF. A number of reactions take place simultaneously in buffer solutions. The chemistry of this mixture is discussed below.

Fluoride and ammonium ions are created from the complete dissociation of the NH_4F salt:

$$NH_4F \rightarrow NH_4^+ + F^-$$

The fluoride can, in turn, react with water and/or the hydronium ion to form HF_2^-.

However, as the conjugate base ($NH_3 \cdot H_2O$) of the ammonium ion, is a weak base, *some* of the NH_4^+ undergoes hydrolysis:

$$NH_4^+ + 2H_2O \rightarrow NH_3H_2O + H_3O^+$$

$$K = \frac{[NH_3H_2O][H_3O^+]}{[NH_4^+]} = \frac{K_W}{K_b}$$

The fluoride and hydrogen (hydronium) ions generated by NH_4F dissociation and hydrolysis, influence the dissociation of the HF present:

$$2HF + H_2O \rightarrow HF_2^- + H_3O^+$$

Thus, the fluoride ion concentration and acidity of the solution are functions of both the HF and NH_4F concentrations. These may be adjusted to control the fluoride and hydronium ion concentrations as well as the fluoride/hydronium ratio. Moreover, these concentrations will remain constant, as undissociated HF acts as a repository for fluoride ions without affecting the etching rate, while the hydrolysis of the ammonium ion stabilizes the pH of the solution.

5.12.3 Uses of Buffered HF

Buffered HF is used in lieu of dilute HF when films, such as photoresist, that could be damaged by a highly acidic environment are present on the wafer surface. Clearly, manipulation of the HF/NH_4F concentrations allows fluoride ion concentrations which are sufficiently high for practical etch rates to be reached, while maintaining acceptable levels of acidity.

BHFs are also useful when a large amount of oxide must be etched. Buffering allows fluoride to be held as undissociated HF and to be released, as free fluoride is consumed. If dilute HF were used, the fluoride concentration would be very high at the start of a process (perhaps so high as to yield an uncontrollable etch rate) and would degrade over the life of the chemical, due to consumption of fluoride and rising acidity depressing further dissociation.

5.13 Hydrofluoric Acid and Metallic Contamination

The effect of metallic contamination on device properties such as gate oxide integrity has been studied and documented by a number of sources [46–48]. Trace metals in hydrofluoric acid solutions have identified as the source of such contamination [49–52].

Elemental silicon in the presence of hydrofluoric acid is a known reducing agent of considerable strength, as demonstrated by the half-cell, with its standard cell potential:

$$Si^0 + 6F^- \rightarrow SiF_6^{=} + 4e^- : \; E_0 = 1.2 \text{ V}$$

If an etching sequence exposes or has exposed bare silicon, then metals can be reduced to their elemental state and deposited on wafer surfaces. This is especially true of noble metals, as the hydrogen ion present in HF solutions is incapable of stabilizing them in their ionic state [49,53].

It should also be noted that the hydrolysis of silicon in basic media is an analogous oxidation reaction. Its half-cell and standard potential are shown below:

$$Si^0 + 6\,OH^- \rightarrow SiO_3^= + 3H_2O + 4e^- :\ E_0 = 1.73\ V$$

This is the reaction that causes the actual surface roughening previously mentioned in the discussion of surface roughening by the SC-1 chemistry. It explains why this phenomenon is such a strong function of pH (note that the half-cell has the stoichiometric requirement for *six* hydroxyl ions). This is why surface etching is reduced when ammonia concentration is lowered.

5.14 Defects Related to Drying

HF processing of Si surfaces leads to a hydrophobic, H-terminated surface, as opposed to oxidizing wet chemistries that leave a chemical oxide terminated by silanol groups. The hydrophobic surface presents difficulties for drying, especially with patterned wafers having oxides and thus mixtures of hydrophilic and hydrophobic regions such that water can be trapped in contact with bare Si during drying. A type of killer defect known as watermarks is formed due to interaction of rinse water with the bare silicon, depending on the effectiveness of the drying step.

The mechanism of watermark formation is commonly understood as follows. Oxygen dissolved in the water (either from its source or diffusing into it from the cleaning tool ambient) reacts with bare Si to form SiO_2, which is subsequently attacked by H_2O resulting in silicates and/or silica dissolved in the water close to the Si surface. As the wafer dries, especially if some evaporation is allowed or water droplets splash onto the surface (for example in spin-drying), these residues precipitate to form stains on the wafer [54,55]. In spin-drying of patterned wafers after HF processing it is virtually impossible to avoid the formation of watermarks. Isopropanol (IPA) based drying techniques have been more successful. One particular method relies on the Marangoni principle [56], wherein a surface tension gradient (STG) created by diffusion of IPA vapor into the water/air interface through which the wafers are drawn leads to a flow of the rinse water from a lower surface tension region to a higher surface tension region. Thus water returns to the bath leaving a dry surface [57,58].

5.15 Polysilicon Etch

This hydrolysis reaction is also responsible for the etching of polysilicon (polysilicon being an amorphous form of elemental silicon). Typically the mixture used is a combination of an alkaline base (KOH) and an alcohol. The reaction itself is an oxidation (as indicated by the half-cell above), so some sort of oxidizing agent is required. In aqueous solution, the hydrogen ion is the predominant oxidizing agent; however, its concentration is suppressed by the basicity of this solution. In the absence of hydrogen, dissolved O_2 must be the oxidizing species. The purpose of the alcohol is to serve as a sacrificial component and modulate the oxygen concentration. As in photoresist stripping, the alcohol is converted in to a carboxylic acid by oxidation. It should be noted that the role of dissolved oxygen is often forgotten. If this process is run in a closed cell processor, where the oxygen content is limited to only that present in the DI water used to formulate the mixture, the reaction (etch) rate drops to zero after the first few lots of wafers are processed.

5.16 Selective Nitride Etch

The use of silicon nitride films in modern semiconductor devices is increasing. Selective etching of silicon nitride (preferentially to silicon oxide) is increasingly important. It is one of the most aggressive and least understood processes in semiconductor manufacturing. Boiling, concentrated, phosphoric acid is the predominant chemical used. Typically, it is held at a constant temperature and concentration by the

addition of deinonized water (as used, phosphoric acid is 85% by weight). So little is known about this process that it is not understood if it is the temperature, the concentration, or the fact that it is boiling that is the controlling factor in this process. The definitive paper on this subject remains, one published in 1967 by two researchers at Bell Laboratories [59]. Key to the process is the formation of pyro-phosphoric acid ($H_4P_2O_7$). The mechanism is conjectured to involve hydrolysis of Si_3N_4 to form hydrated silica and ammonium phosphate. The acidity of the solution provides selectivity by preventing hydration of the Si–O bonds in areas covered by SiO_2. In studies of phosphoric acid etching at Tohoku University, ammonium phosphate was identified as one of the species present in a phosphoric acid etching solution being studied.

5.17 Oxide/Nitride Etch

Some device designs require the etching of silicon oxides and nitrides at equivalent rates. Mixtures of hydrofluoric acid in a polar organic solvent (e.g., glycerol or ethylene glycol) are typically used for this process. The reaction is run at elevated temperature (70°C–80°C) to give an acceptable etch rate.

Etch mechanisms of Si_3N_4 and SiO_2 in HF solutions are discussed in detail in reference [60].

5.18 Bulk Organic Removal/Photoresist Strip

Photoresist that remains on wafers following lithographic processes is typically a highly cross-linked organic polymer that is very difficult to remove. In the early days of semiconductor manufacturing, these were removed with liquid chemical stripping agents–typically, either mixtures of highly aggressive organic solvents or mixtures of sulfuric and nitric acids. The industry has moved away from the use of organics due to their toxicological and environmental impact (one of the more popular stripping agents was a mixture of phenol, benzylsulfonic acid and chlorobenzene in a solvent of perchlorethylene). However, sulfuric/oxidizer mixtures continue to be used as post-photo cleans and/or heavy organic strips.

5.19 Photolithography

The photolithographic process is similar to the photographic printing process. Light is projected through a mask (analogous to a photographic negative). The wafer is covered with a photo-active resist (photoresist—analogous to photographic emulsion). Where light transmitted through the mask strikes the photoresist, it causes a chemical reaction in a compound that is intended to cross-link the resist polymer (in the case of positive resists, light breaks down the cross-linking agent and renders it inactive; while in negative resists, the agent is activated by light). In subsequent steps, the image is developed and stabilized, including a bake to remove residual solvents. The cured photoresist forms a protective mask to protect areas of the wafer surface from subsequent processing steps such as etching or implantation. As a result of these treatments, the polymer can be thought of as "work hardened" and is usually very difficult to remove.

Today, plasma ashing is used to remove the bulk of the resist, but the ash process is often stopped short of removing the entire photoresist film. Photoresists can contain significant levels of metallic contaminants. If photoresists are ashed down to the wafer surface, some of these metals can become attached to the wafer surface. For this reason, the ashing process is stopped short of the surface, and wet stripping is used to remove the remaining resist and to solvate residual metals.

5.20 Sulfuric/Oxidizer Chemistry

The first photoresist stripping solution, 9:1 sulfuric/nitric acid, was not an invention of the semiconductor industry. Rather, it is a common nitrating agent used in synthetic organic chemistry [61]. Nitrates are famous in the field of synthetic organic chemistry due to their commercial importance

for two important properties: (1) they are highly reactive and form the basis for many explosives and (2) they have high solubility and ability to form colloidal suspensions in water. Cellulose nitrate was one of the first polymers in commercial use, as sheets could be cast from aqueous colloidal suspensions (the first photographic film used cellulose nitrate as a base). In photoresist stripping, nitration is a successful means of organic removal, as the nitration reaction accomplishes cission of long photoresist polymer chains (breaking them into shorter chains) as well as nitration of the polymer fragments to render them water soluble. One of the more curious aspects of photresist stripping is that, despite the high organic loading from repeated use of the strip bath on multiple wafer lots, the bath itself remains relatively uncontaminated. While the sulfuric/nitric mixture renders the organic water soluble, at 96% by weight sulfuric acid and 70% by weight nitric acid, the bath itself is essentially anhydrous. Solvation (removal) of the photoresist takes place primarily in the rinse bath following the sulfuric/nitric tank. A word must be said concerning the ability of sulfuric/nitric solutions to form explosive organic compounds. Due to the solvation mechanism discussed above, an explosive concentration of organic nitrates is not reached in the sulfuric/nitric bath. The organic nitrate concentration in a constantly replenished rinse tank also remains low. However, sulfuric/nitric systems are typically used at elevated temperature. A bath using an organic heat transfer fluid for heating is potentially dangerous, as a pinhole could develop in the vessel containing the acid. Leakage of sulfuric/nitric into the heating jacket could result in the nitration of the heat transfer fluid over time.

5.21 Reaction Mechanism

In general, the reaction mechanism for an organic (R) with sulfuric/nitric acids can be written [61]:

$$HONO_2 + 2H_2SO_4 \rightleftharpoons H_3O^+ + 2HSO_4^- + NO_2^+$$

$$NO_2^+ + RH \rightarrow R^{\oplus}(H)(NO_2)$$

$$R^{\oplus}(H) + HSO_4^- \rightarrow RNO_2 + H_2SO_4$$

The action of sulfuric acid should be carefully noted in this reaction sequence. The tendency of sulfuric acid to dissociate and form the hydronium ion (H_3O_+) is so strong as to force nitric acid (usually a strong acid) to act as an Arrhenius base in this reaction. This first step is a required part of the reaction sequence to form the nitronium ion (NO_2^+). It only takes place if the reaction environment is essentially anhydrous. If free water were present, the first ionization of sulfuric acid and formation of the hydronium ion would take place with water (as below), not nitric acid, and the nitronium ion would not be formed:

$$H_2SO_4 + H_2O \rightarrow H_3O^+ + HSO_4^-$$

In an overall examination of this reaction sequence, sulfuric acid serves as a dehydrating agent—removing an OH^- group from HNO_3 and a hydrogen from the organic molecule. Unless the solution is kept anhydrous, this reaction sequence ceases to function.

5.22 Sulfuric/Peroxide (Piranha)

Today, hydrogen peroxide has by-and-large supplanted the use of nitric acid in this class of cleans. The reaction sequence is similar, with the exception that oxidation continues until a water-soluble carboxylic acid (RCOOH) is formed.

Although the substitution of hydrogen peroxide for nitric acid offers a number of advantages, its use makes maintenance of anhydrous processing conditions more difficult. Hydrogen peroxide is an unstable material and continuously undergoes the dispropotionation reaction:

$$2H_2O_2\ \tau\ 2H_2O + O_2\mu$$

Thus, the use of hydrogen peroxide results in the continuous dilution of a piranha bath. For this reason, piranha dip tank systems are typically limited to 2–4 h of bath life before they are sufficently dilute so as to be unable to strip resist. Re-addition (spiking) of fresh hydrogen peroxide is not an effective means of regeneration, as the typical hydrogen peroxide solution used in semiconductor manufacturing is 30% (by weight) peroxide. The remaining 70% is water, so any attempt to spike a bath results in its further dilution.

Due to the small chemical volumes used and the high heat of solution evolved when hydrogen peroxide and sulfuric acid are mixed in-situ, spray processing is an economic and effective means of carrying out piranha processing.

5.23 DI/Ozone

With a standard half-cell potential of 2.07 V (vs. 1.776 V) in an acidic medium, ozone (O_3) is a much stronger oxidizing agent than is hydrogen peroxide. Rather than stopping at the formation of an intermediate nitrate or a carboxylic acid, ozone possesses sufficient oxidizing strength to decompose organics fully to CO_2 and water. Sulfuric acid is no longer necessary as a dehydrating agent, and the process takes place in a deionized water medium.

Increasing the efficacy and rate of reaction has involved increasing the concentration of ozone in solution. Traditionally, photoresist stripping had been a high temperature operation; however, it was noted that the solubility of ozone in DI water (as with most gasses) *increases* with *decreasing* temperature. Ozone injected into chilled DI water has been shown to be effective in both increasing the number of types of resist removed and/or increasing the stripping rate of resists from wafer surfaces [62]. While cooling increases the saturation concentration of ozone, decreasing the solution temperature has the deleterious effect of decreasing the activity of the reactant (oxidizing strength). In the Nernst relationship, oxidizing potential is directly proportional to the absolute temperature at which the reaction takes place, while it is only logarithmically-dependent upon the concentration of the reactants. Other techniques used to increase ozone solubility include injecting it into water in a pressurized vessel [63].

Unlike most semiconductor wet processes, the DI/Ozone process is heterogeneous (has reactants in two phases). As a result, the contacting method plays a major role in the reactivity of DI/Ozone mixtures. Several equipment manufacturers use off-line mixing, while others inject ozone gas directly into a processing chamber, so gaseous ozone is present in addition to that dissolved into the processing liquid. Wafers are either immersed in a bath of the processing fluid or are sprayed with a film of the liquid solution [64]. Another technique uses ozone gas and hot water vapor in a slightly pressurized chamber [65].

Because of the silicon loss issue in the source and drain extension regions around the gate, there is recently increased interest in non-plasma methods for stripping photoresist after the ion implantation steps that dope the Si. However, the high implant doses (often in excess of 1E15 atoms/cm_2) polymerize the photoresist the photoresist and makes it more difficult to remove. Ozone water processes have so far

proven incapable of effectively removing such high dose implanted photoresist, and high temperature piranha type processes appear more promising for this application.

5.24 Surface Preparation and Cleaning for Interconnect

Back end of line surface preparation challenges are similar and yet different from what is experienced in the FEOL. Concerns with defect density, etch rates, etch uniformity, silicon, and dielectric consumption, PRE, damage to fragile structures, surface contamination, throughput, edge of wafer/bevel cleaning, backside cleaning and drying are all common. The difference is driven by which process and integration parameters are taken into account when determining the optimum conditions. Additional concerns include metal corrosion, photo diode induced copper re deposition (PICR) [66,67], film adhesion, modifying the dielectric constant (k) of the low-k dielectric, drying when porous films were exposed to liquids, and fundamental mechanical strength of 7+layers of interconnect (some of the advanced microprocessors now use up to 11 layers of interconnect). Surface preparation interacts in new ways with each of these issues creating, significant challenges.

When discussing surface preparation, one cannot fully understand the relative merits or applicability of a clean without placing the clean within the context of the challenge. In the following sections for subtractive aluminum interconnect and copper/low-k dual damascene interconnect, the fundamentals of the process integration scheme for each will be outlined before proceeding to discussion of the specific cleaning challenges. A detailed discussion of this topic can be found by B. K. Kirkpatrick at [68].

5.25 Subtractive Aluminum Interconnect

To form a single interconnect layer one must build two sub-layers, contacts or vias and a metal stack. Contacts and vias are typically formed using a tungsten (W) plug process. W-plugs are created by depositing a dielectric, plasma etching a dielectric hole (the contact or via), performing a post etch clean (dry+wet), placing a barrier material down into this hole using physical vapor deposition (PVD), filling the hole with W by chemical vapor deposition (CVD), and then polishing the top surface off using chemical mechanical polish (CMP) and closing with a post CMP clean. The CMP step serves both to separate the filled holes and to planarize the surface. This same integration loop is typically used for contacts regardless of whether the following integration utilizes subtractive aluminum or Cu dual damascene. For subtractive aluminum integration schemes, the W-Plug sub-layer is reused after contact between each metal layer. In some instances, the top most vias are so large that the metal stack itself can be used to directly contact the underlying metal layer.

Subtractive aluminum metal stacks are formed on top of contacts or vias and are made by the following sequence: deposit a metal film, photo pattern the metal film, plasma etch the metal film, perform a post etch clean (dry+wet), and then, start over with a via layer. The metal film is typically Al-(0.5%–2.0%) Cu sandwiched between TiN, Ti/TiN, or Ti, with either organic or inorganic bottom anti-reflective coating (BARC) on top. In some instances, W or WSix is used as the first metal layer. Due to resistivity requirements, this is more of an option for Memory devices and is rarely seen with Logic devices. Table 5.7 provides additional detail on subtractive aluminum processing with tungsten plugs.

As described above, three surface preparation steps are required at a minimum for each interconnect level. These are, post contact/via etch clean, post CMP clean, and post metal stack etch clean. Chemical mechanical polish and CMP cleans are described in detail in another chapter and therefore will not be discussed herein. Figure 5.6 shows both the typical metal stacks as well as the cross-sections of W-plugged contacts [68].

TABLE 5.7 Typical Subtractive Aluminum Process Flow Using Tungsten Plugs

	Subtractive Aluminum with Tungsten Plugs	
Step	Description	Purpose
1	Existing underlying metal level	Layer to electrically contact
2	Deposit an etch stop and inter-level dielectric (ILD)	Insulator between vias (laterally) and metal layers (vertically)
3	Via pattern and etch	Define and open up via holes
4	Post via etch clean	Removes the post etch polymers
5	Deposit metal barrier	Prevents metal from diffusing throughout the dielectric
6	Deposit tungsten	Fills the via hole
7	Chemical mechanical polish (CMP) excess off	Removes all tungsten not contained within the via
8	Post CMP clean	Removes residual polish defects; slurry residues, particles, etceteras
9	Deposit a metal stack (multi-step typically including a barrier, aluminum alloy, and top anti-reflecting coating)	Electrical conductor
10	Clean surface	Clean surface prior to patterning
11	Pattern metal	Define areas on metal to leave behind, these will become the electrical wires
12	Plasma etch pattern through metal stack stopping on underlayer	Remove all areas of metal not covered by the pattern
13	Clean post etch polymers	Remove defects from the top and sidewall of the metal wires

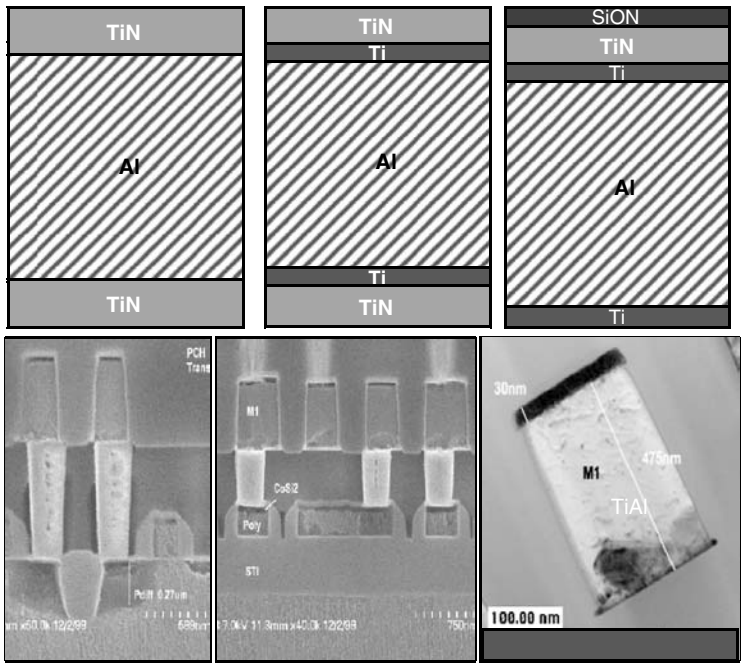


FIGURE 5.6 Typical Al/0.5% Cu metal stacks along with cross-sections.

5.26 Key Subtractive Aluminum Cleaning Challenges with Associated Defects

Post plasma etch cleans for both the metal stack and via are combinations of ash, followed by solvents and/or dilute acids, sometimes followed by another ash. In some specific cases, the wet portion is a low temperature, acid mixture. For example, dilute sulfuric peroxide plus hydrofluoric acid mixtures at 25°C has been reported by Archer and Coteau [69]. For the most part, ash is dominated by oxidizing chemistries [70,71]. Besides oxygen, additional gases such as nitrogen (as a diluent), forming gas (3% H_2/N_2), and fluorocarbons are sometimes used. In most cases, the wet portion of the surface cleaning is performed utilizing an etch mechanism. In other words, the top surface of the film is etched off (Al–0.5% Cu or dielectric), taking with it the defects.

This leads to slightly higher line resistance but fewer shorts and generally higher yield. From the via perspective, using an etch cleaning mechanism drives via size, to increase with resulting decrease in via resistance. Negative impact on yield from via to unrelated metal shorting is not typically seen given the metal design rules used. In fact, contact, and via cleans utilizing etch cleaning mechanisms usually result in higher yields. The key statement being, "given the metal design rules used." With the exception of some memory devices that do not tend to drive high interconnect density, the majority of all new process designs utilize copper/low-*k* dual damascene interconnect.

5.27 Copper/Low-*k* Dual Damascene Interconnect

Copper dual damascene can be implemented in several ways. The four most common integration schemes are self-aligned, partial (or half) via first, via (or full) first, and trench first. Of these, the most common is full via first (Further comparison of these processes is included in another chapter). All of these schemes utilize a stack with a via etch stop that doubles as a Cu diffusion barrier and some sort of inter-level dielectric/intra-level dielectric (ILD/IMD) stack. Within this ILD/IMD stack there may or may not be a trench etch stop and a top hardmask layer. In all cases, there are at least three cleans (Figure 5.7, [68]) required to complete one interconnect level, post via plasma etch, post trench etch with integrated etch stop clean and post Cu CMP. Each of these cleans may be a single or a multi-step clean sequence with at least one of these steps being wet. In some cases, no clean is performed until the etch stop has been cleared (step A in Figure 5.7 would be eliminated). Table 5.8 contains additional data on copper dual damascene using the full via first approach.

5.28 Key Cu/Low-*k* Challenges with Associated Defects

Cu/low-*k* dual damascene requires fundamental change to past cleaning strategies. Besides the standard cleaning requirement of minimizing defectivity, there are many new challenges and some old challenges

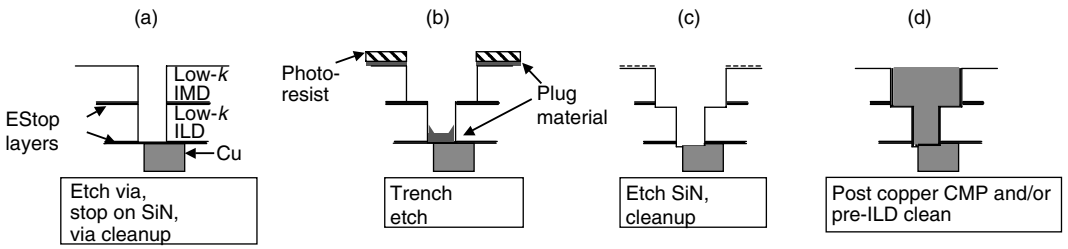


FIGURE 5.7 Cleaning needs with a via-first, Cu/low-*k* dual damascene integration scheme.

TABLE 5.8 Typical Copper Dual Damascene Flow Using a Full Via First Integration Scheme

	Copper Dual Damascene	
Step	Description	Purpose
1	Existing underlying metal level	Layer to electrically contact
2	Deposit an etch stop and inter-level dielectric (ILD)	Insulator between vias (laterally) and metal layers (vertically)
3	Deposit an etch stop and intra-level dielectric (IMD)	Insulator between metal lines (laterally)
4	Via Pattern and etch	Define vertical connections between layers
5	Post via etch clean	Removes the post etch polymers
6	Trench pattern and etch	Define horizontal (lateral) connections within a layer
7	Post trench etch clean	Remove the post etch polymers
8	Etch the via etch stop	Remove the etch stop exposing the underlying copper wire
9	Post via etch stop clean	Remove the post etch polymers with high selectivity to metal
10	Deposit metal barrier and seed	Prevents metal from diffusing throughout the dielectric and provides a seed layer required by electro chemical deposition (ECD)
11	Deposit metal (typically ECD copper)	Deposits bulk metal to fully fill the via and trench
12	Chemical mechanical polish (CMP) excess off	Removes all copper from the top wafer surface not contained within a via or trench
13	Post CMP clean	Removes residual polish defects; slurry residues, particles, etcetera

requiring very different solutions. In the following paragraphs, the authors will attempt to address the key challenges and associated defects in no specific order.

Copper does not self-passivate as Al–0.5% Cu does. While aluminum builds up a passivating oxide, copper does not. Copper oxide is of a poor quality and does not become sufficiently dense to become a diffusion barrier capable of blocking further oxidation. This results in continuing oxidation for as long as copper is exposed to an oxidizing ambient and can result in the complete oxidation of all copper present. Unfortunately, where aluminum and copper differ in the ability to self-passivate, the oxides formed from both metals share the property of being poor conductors. Since the Aluminum oxide is relatively thin, for subtractive aluminum metallization this can be largely ignored. That is not the case for copper metallization schemes. Copper cleans require a strategy to minimize the amount of oxidized copper remaining after the clean is complete. Luckily, multiple solutions exist. These include: minimize the total duration and frequency upon which Cu is exposed to cleans; use reducing chemistries in place of oxidizing chemistries [72]; use reducing chemistries after oxidizing chemistries [72,73]; remove oxidized Cu with wet chemistries that are selective to Cu while maintaining high etch rates on copper oxide [74,75]; or physically remove oxidized Cu with a pre-sputter etch (pre-PVD) clean [76]. The best solution is largely dependant on the integration scheme used and the equipment and chemistries already available in the wafer fab.

Besides oxidation corrosion, non-passivated metals can be susceptible to PICR [66,67]. Whenever metal lines are in close proximity, are connected to a sufficiently large silicon photo-diode, and are exposed to visible light (the band-gap of Si is 1 eV, all visible light exceeds 1 eV energy level) the possibility for PICR exists. This is largely unseen with aluminum metallization schemes since aluminum has such a strong tendency to self-passivate. But, as previously discussed, copper does not self-passivate. Figure 5.8 illustrates how Cu^{2+} can drift and/or diffuse away from a Cu anode redepositing on a cathode [68]. Since the electric-field lines in the dielectric are primarily perpendicular to the Cu^{2+}, the drift component is negligible. Where the same is largely true for the diffusion component in a dry environment, diffusion can be significant in a wet environment. Figure 5.9 shows top down scanning

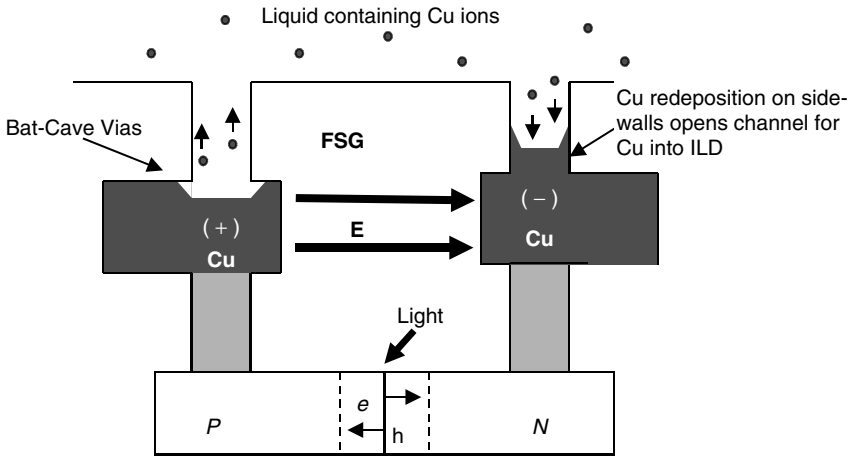


FIGURE 5.8 Photodiode induced copper redeposition (PICR) inside a via.

electron microscopy (SEM) pictures with Cu-leads connected across a large photo-diode. While several solutions exist for this challenge the simplest and lowest cost is to perform all wet processing of copper in the dark preventing light from reaching the photo-diode. Of course several of the process steps in the BEOL do not have exposed copper and correspondingly, one may propose to use non-darkened tools for these steps. However, by taking this approach, if there are other process excursions copper may still be exposed, resulting in damage from PICR.

Barrier integrity plays an important role in preventing metal migration to transistors. Good barriers do not exist without good surface preparation. Any process using a strong physical component (such as plasma etch, ash, pre-sputter etch, high density plasma processes (HDP), …), tends to sputter exposed films (suspect processes almost always use an RF-bias) [77]. Once these other processes have been optimized to minimize the movement of metals to undesired locations, the surface preparation challenge is to remove the relocated metal if it will be beyond where the barrier will be formed. For subtractive Al processing, the integration scheme and cleans relying primarily on etch cleaning mechanisms tended to alleviate this problem. As the top surface was etched away, defects including relocated metal tended to be removed as the surface was etched. For Cu low-*k* BEOL, the same cannot be said. Cu vias and lines are formed within the already etched holes or trenches in the dielectric. In many cases, these features approach the minimum resolution of the available photolithography equipment. Thus, cleans using any appreciable etch cleaning mechanism increases the probability for electrical shorts to dissimilar metal.

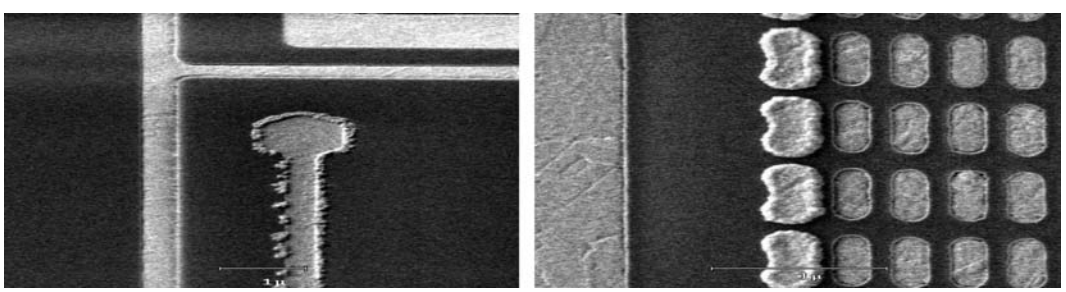


FIGURE 5.9 Photodiode induced copper re deposition (PICR) example. Cu trenches contacting *N*-well, photo diode turned on, cathodes for PICR.

The surface preparation challenge is to remove Cu containing residues from the feature sidewalls while preserving critical dimensions (CDs) and any exposed Cu at the bottom.

Another aspect of barrier integrity is cross-contamination. A barrier film is useless if lax Fab protocols allow metals or other contaminants to be accidentally deposited beyond the barrier. Cross-contamination has been an on-going challenge for years. Multiple Fabs that have been completely shut down by contamination incidents resulting in thousands of scrapped wafers, numerous tool wet cleans, replacing most if not all of the Fab's quartzware, many-many hours of downtime and millions of dollars in lost revenues. In the past these incidents were typically related to accidental contamination from unknown metals or sodium (Na) contamination. With the advent of Cu damascene (Cu has a higher diffusion constant than Na), Cu provides another serious contamination risk. Solutions must take into account both prevention and mitigation. Prevention takes the form of strict limitations for equipment usage between Cu portions of the flow and non-Cu portions. This includes Al since most Fabs have not implemented barriers sufficient to prevent Cu penetration on their Al based flows. Rigorous cross-contamination testing should be performed on any tools that are shared [78]. Mitigation most often takes the form of cleans targeted exclusively at the wafer backside [79]. Backside copper removal cleans enable additional tool sharing. This significantly eases restrictions on many low temperature tools such as metrology and photolithography.

Interface adhesion of new materials has always been a challenge and is likely to remain so. The difference with Cu/low-k dual damascene is the number of new materials (new is relative to the 35 years experience the industry has with subtractive Al processing), the generally reduced adhesion of almost all low-k materials, and the increased number of interfaces involved. Other process areas tend to be primarily responsible for basic material properties. Surface preparation can impact the situation by controlling the frequency, duration, and chemical selectivity of those cleans exposed to the interfaces in question. Figure 5.10 shows top down SEM and cross-section TEM photos of a SiN barrier delaminating off the top of a large Cu area. By controlling the chemistry, duration, and frequency of cleaning chemistry exposure, delamination was eliminated. Delamination tends to be dominated by lateral etching along interfaces. Therefore, simply decreasing total exposure time to the surface preparation chemistry nearly always decreases the probability a film will delaminate.

The importance of fully drying the wafers is more important than ever. The low-k films currently in production fall into the low-k, very-low-k and ultra-low-k categories (LK, $4.2 \geq k \geq 3.6$; VLK, $3.6 \geq k \geq 2.9$; ULK, $k < 2.9$). From VLK dielectrics onward, porosity is a processing factor. While dielectric suppliers are developing materials with closed pores, most acknowledge some percentage of open pores. To further increase the difficulty, many of these ULK materials have organic functional groups such as methyl added to reduce k. Besides reducing k, methyl groups cause the surface to be hydrophobic. If some sort of plasma processing on the top surface is not performed prior to a wet clean, conventional spin dry or boiling sump IPA vapor dryers result in film defects (watermarks, k value shift, particles, …). For example, Figure 5.11 shows a Fourier transform infrared spectroscopy (FTIR) spectra of wafers treated with a reducing ash, then various wet chemistries followed by commercially available dryer [68]. As can be seen, the silanol peak (Si–OH) around 3400 wave numbers is still noticeable. The addition of water to these films increases the k values.

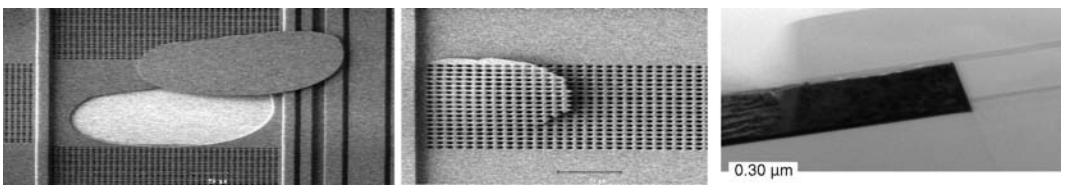


FIGURE 5.10 Top down scanning electron microscopys (SEMs) and cross-section TEM showing interface adhesion problems on a Cu/low-k device. Barrier delaminating off the top of Cu.

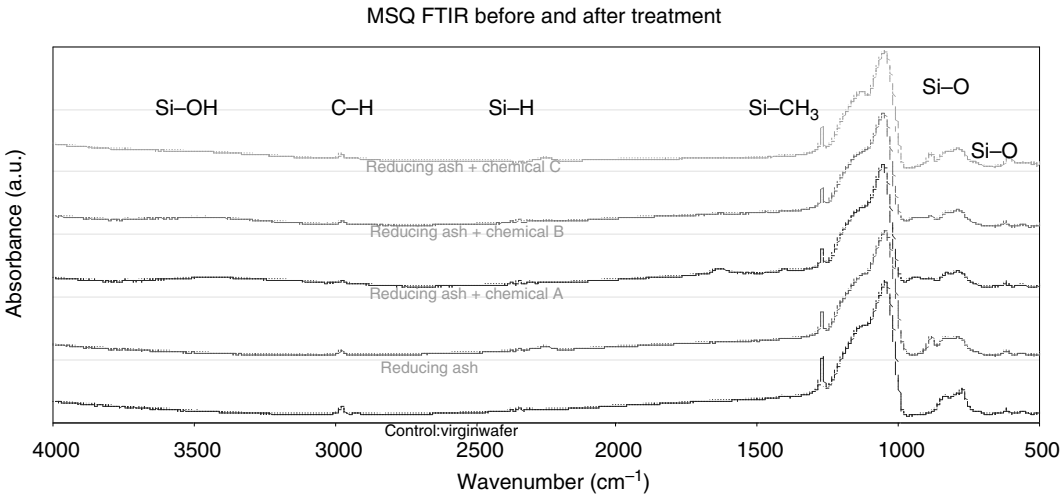


FIGURE 5.11 Porous methylsilsesquioxane (MSQ) film change vs. cleans by Fourier transform infrared spectroscopy (FTIR). Significant film change apparent vs. clean treatment.

To address the efficiency of drying with the proliferation of hydrophobic or partially hydrophobic surfaces, new wafer drying techniques were introduced. Some of these include Marangoni [80,81], Rotagoni [82,83], STG, [84,85] and direct vapor displacement [86,87]. The one common theme for all of these methods is to utilize the difference in surface tension between dionized water (DIW) and a drying agent (typically IPA) to allow the drying agent to physically displace the DIW.

Super critical fluids (SCF) cleaning and drying is another novel approach developed to address some of these same issues arising from cleaning and drying hydrophobic, porous low-*k* films [88–91]. In this case, a wafer is placed into a chamber with a gas at elevated pressure and modest temperature ($\geq$1070 psi, 31°C for CO_2) such that it becomes a SCF. Super critical fluids have viscosity and diffusivity like a gas, nearly zero surface tension, and a density like a liquid. This allows a SCF to flow freely through nearly any pore. Add to this fluid a co-solvent that has the property of increasing the cleaning efficiency and you have a cleaning solution capable of cleaning and drying without capillary force, leaving no liquid behind to change the dielectrics *k* value and inherently avoiding water marks. These properties make a SCF uniquely capable of dealing with open pores in ULK films. However, due to relatively high cost per wafer and significant environmental and technical challenges brought on by recovery and purification of high volumes of CO_2, widespread adoption of this surface preparation technology is unlikely until such time as it becomes enabling. At this time, these authors do not find this likely until such time as open pore ULK films are introduced, if then.

Critical dimension control is another ongoing challenge. For the most advanced technologies, typically the first few layers of interconnect (signal leads) are designed at minimum photo capability. In the past, most wet portions of the surface cleaning were performed utilizing an etch mechanism. For subtractive aluminum metallization, this drives slightly higher line resistance but fewer shorts and generally higher yield. For damascene processes, it tends to work in exactly the opposite way. Etch cleaning mechanisms increase metal and via CDs, decreasing the distance to unrelated metal. This both increases capacitance's decreasing performance, and decreases yield from increased via chain and comb-serpent shorting. Because of this, BEOL cleans tend to focus on more highly selective etch chemistries (selective to the underlying films) and clean mechanisms that modulate zeta potential or solvation.

Dielectric damage and the resulting increased material loss during wet cleans has been a concern for some time for BEOL processing. Like many parameters, the acceptable amount of damage has been

continuously shrinking with each new technology node. Most of the new low-k materials utilize doping additions of fluorine, carbon and/or the addition of methyl groups ($-CH_3$) to decrease k values. Unfortunately, many etch and ash processes tend to remove these additions from the surface of the material degrading electrical properties. Besides increasing k values and leakage, the removal also results in an enhanced wet etch rate. Control of this surface change is critical to allow acceptable control of the resulting dielectric constant, capacitance, and CD.

When considering film loss, process interactions are always an important concern. For example, film loss in a wet chemistry may be increased by orders of magnitude depending on previous processing. In the example shown in Table 5.9, the author compares FSG film ($k=3.6$) and organo silicate glass (OSG) film ($k=2.9$) loss caused by five min exposure to various commercially available wet cleans, to the same wet cleans following an oxidizing or reducing chemistry ash. As one can see, the plasma ash process creates damage that resulted in increased film loss in the following wet clean by more than an order of magnitude.

5.29 Typical Defects

The last challenge to be discussed herein is defect density including what constitutes a killer defect and what are the typical defects. Please note, this will not be an exhaustive treatment on this subject as there is an entire chapter dedicated to this subject later in this book. With subtractive Al processing, a defect landing on the surface could immediately be classified as killer or benign. For the most part, the defect either shorted the metal to unrelated metal or it did not. With damascene processing, this is not the case [92–94]. Defects can still short together unrelated metal. Defects can also land within the via or trench subsequently displacing the metal. If enough metal is displaced, the line becomes an open circuit and the defect can be detected and eliminated from the product distribution at end of line electrical test. But, if the defect is not sufficiently large to open circuit the line at time zero, it often becomes the site, for early in the field failures through an electromigration fail mechanism [95–97]. If at time zero, surface particles do not create shorts or opens, these particles can still be encapsulated by the dielectric. The encapsulated particle can then interact with the CMP steps possibly leading to rip-out defects. Figure 5.12 tracks both a particle and a micro-scratch through the process flow illustrating how these become shorts or "killer" defects [68].

Cu/low-k dual damascene can have many different kinds of defects. Eight of the most common types as detected after end of line physical failure analysis are shown in Figure 5.13. These include via voiding/electro chemical deposition (ECD) voids or floating vias, missing vias, metal filaments, missing metal, blocked trenches, scratches, interlevel dielectric ripouts, and pre-metal dielectric (PMD) particles. While the majority of these defects result in open circuits (Figure 5.13a through c, e, f, and i), a few create short circuits (Figure 5.13d) while others can create either an open or a short circuit (Figure 5.13g and h). Figure 5.14 shows six of the most common type of defects as detected inline. These include large flakes,

TABLE 5.9 Comparison of Wet Etch Rates in Various Commercially Available Strip Chemistries vs. Starting Conditions

	Angstroms of Substrate Etched in 5 min					
	Chem A	Chem B	Chem AC	Chem D	Chem E	Chem F
FSG no ash	3	19	5	25	<1	58
FSG ox	42	84	64	93	<1	62
FSG red	36	63	31	69	<1	38
USG no ash	<1	17	1	23	<1	18
USG ox	30	67	62	73	4	65
USG red	28	36	23	41	2	32

Starting conditions include as-deposited, following a one minute oxidizing ash and a one minute reducing chemistry ash.

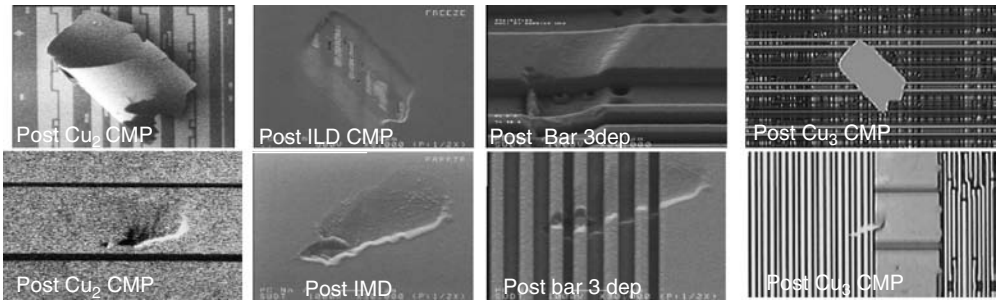


FIGURE 5.12 Partitioning test from M2 to M3 chemical mechanical polish (CMP) showing how particles (top) and scratches (bottom) can cause shorts. (From Kirkpatrick, B. K. Surface Preparation Challenges with Cu/Low-*k* Damascene Structures. ECS Fall, 2001.)

small particles, voids, pits, polymer strands (post etch residues), and ripouts. All of these defects have potential surface preparation mechanisms.

For ECD voids, one mechanism is to have incomplete polymer strips. These residual polymers then block or shadow the PVD barrier/seed deposition. Without a seed, the Cu will not plate into that feature. Missing contact, via, or trench patterns can have many causes. Surface preparation itself may impact missing patterns multiple ways. These include leaving surface particles or flakes, leaving chemical residues that later poison deep ultraviolet (DUV) photoresist, or use of nitrogen in the dry-clean plasma that can lead to later amine out-gassing and poisoning of deep ultraviolet (DUV) photoresist. Even the age old problem with watermarks (drying residues on hydrophobic surfaces) can lead to missing photo patterns. Inter-level dielectric/metal ripouts have already been discussed above. Floating or incomplete

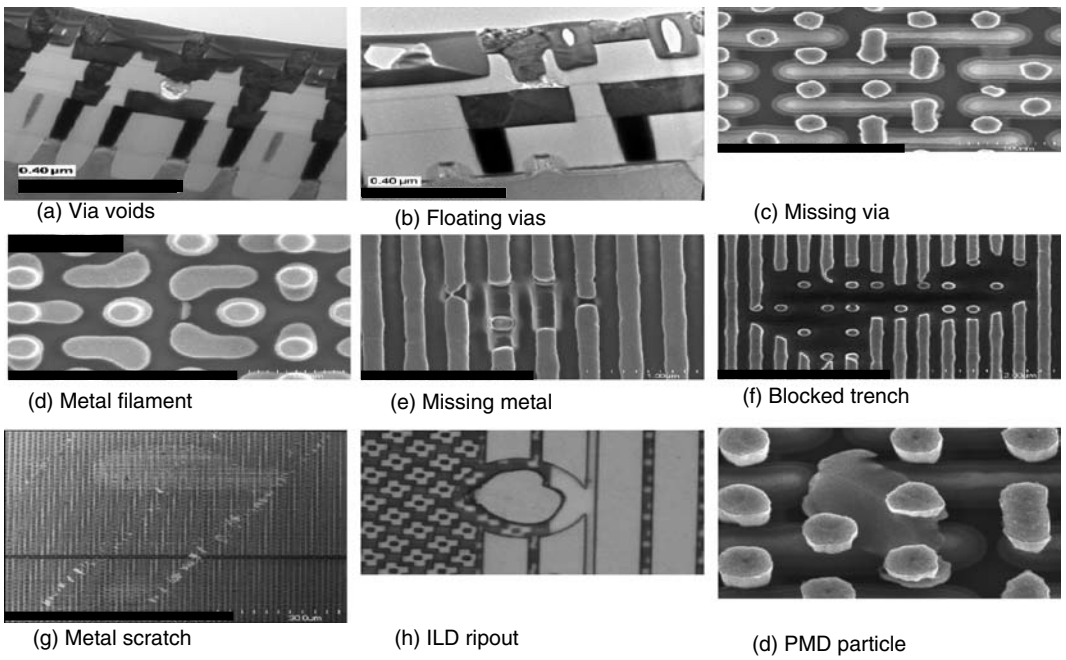


FIGURE 5.13 Nine of the most common Cu/low-*k* dual damascene defect types as detected after end of line physical failure analysis.

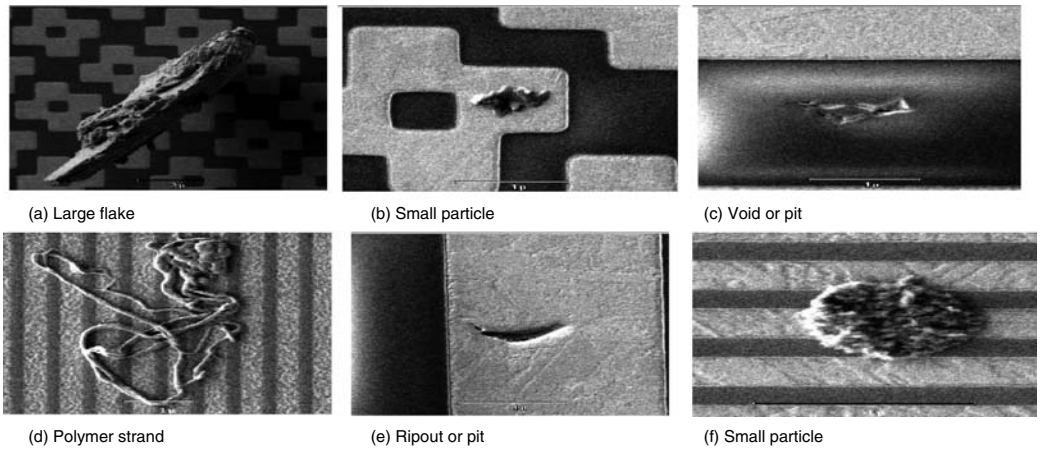


FIGURE 5.14 Six of the most common Cu/low-*k* dual damascene defect types as detected inline.

vias represent another one of the common defects that can have many causes. From a surface preparation perspective, floating vias are most likely due to some sort of blockage. Besides the particle, flake and poisoning mechanisms discussed already, simple incomplete residue removal from a previous post etch clean can create this defect.

5.30 Control and Monitoring of Surface Treatment Processes

As with any semiconductor manufacturing step, control of surface cleaning and etching processes is critical. For example, the importance of maintaining proper chemical ratios in SC-1 baths has been stressed. A number of analytical control techniques and monitoring strategies are available to the process engineer. It is customary in manufacturing to monitor at least particulate addition to wafers, as well as etch rates and uniformity, on a regular basis. In many cases, metallic contamination in baths is also monitored. A list of common analytical techniques is provided in Table 5.10.

TABLE 5.10 Common Analytical Techniques Used in Monitoring and Characterization of Surface Treatment Processes

Parameter	Technique	Notes
Particles added	Laser surface scanner	Typical daily check
Particle characterizing	Scanning electron microscopy (SEM)/energy dispersive x-ray spectroscopy (EDX)	Energy dispersive x-ray spectroscopy provides elemental analysis. May be used to give spatial distribution of particles of differing compositions
Particles in bath	In-situ laser liquid particle counter	Available for some chemistries
Metal contaminants	Total reflection x-ray fluorescence spectroscopy (TXRF)	Detects common metals (except Al) on wafer surface. May be used to give some spatial distribution of contaminants on wafer surfaces
Metal contaminants	Vapor phase decomposition (VPD)–inductively coupled plasma mass spectrometry (ICPMS) or TXRF	Lower detection limit than TXRF. Spatial distribution not possible
Metal contaminants	Elymat (diffusion length)	Primarily for Fe. May be used to give spatial distribution of impurities on wafer surface
Etch rate/uniformity	Ellipsometer	Film thickness by light scattering
Organic contamination	Thermal desorption system (TDS), time-of-flight secondary ion mass spectrometry (TOF-SIMS)	Analysis of surface organics

Location and quantification of particles on a surface is accomplished with laser surface scanners, which utilize light scattering to detect disruptions in a smooth, polished silicon surface. Current particle size detection limits for such equipment are approximately 90 nm. In addition to Si or SiO_2 monitor wafers, laser surface scanners are available that can scan patterned wafers. For further characterization of contaminants, SEM can be used in conjunction with energy dispersive x-ray spectroscopy (EDX) [98].

The most common industry technique for analyzing metal contaminants is total reflection x-ray fluorescence spectroscopy (TXRF). X-rays impinge at a very low angle on the sample surface are reflected, exciting only the top few monolayers. Fluorescence x-rays are subsequently emitted from these monolayers in various directions, and collected by a detector. The detector is an energy dispersive spectrometer which analyzes the x-rays according to energy, giving elemental information [99]. Generally, elements of atomic number less than Si or greater than tungsten (W) are not detected. Only a small area of the wafer surface is analyzed. Detection limits for TXRF vary for different metals, and are mostly on the order of 10^{10} atoms/cm^2. The limits can be significantly lowered if vapor phase decomposition–droplet collection (VPD–DC) is used [100]. With this technique, the native or chemical oxide on the Si wafer surface is dissolved using vapor HF. Impurities which had been in or on the oxide will remain in the resulting film of dissolved oxide on the wafer. A small amount of collecting liquid (normally HF based) is drawn across the wafer surface, and metal impurities concentrated in this liquid are then analyzed using TXRF or inductively coupled plasma mass spectrometry (ICPMS). Resulting detection limits can be well below 10^9 atoms/cm^2 [100]. While it is a useful technique for extending the sensitivity of TXRF, this methodology has some drawbacks. As TXRF samples a small area of the wafer, concentration distributions can be determined using TXRF. This information is lost when using VPC-DC. Metals not complexed by the fluoride ion may be lost, and, if bare silicon is exposed as a result of surface etching, noble metals may be lost by silicon reduction.

Organic contaminants can be detected using time-of-flight secondary ion mass spectrometry (TOF-SIMS), in which ionized surface atoms are ejected by impact from a beam of $\sim$1 keV ions and subsequently separated according to their mass [101]. This determines elemental composition of surface layers. Another technique is thermal desorption spectroscopy [102]. This technique is based on the fact that adsorbed atomic or molecular contaminants will desorb according to the strength of their bonds. Samples are subjected to a constant-rate temperature ramping, and desorbed components are analyzed using a mass spectrometer or gas chromatograph. These techniques identify fragments of organic species. A technique for quantifying actual carbon atoms in gate oxides grown after cleaning processes, using SIMS depth profiling through a polysilicon encapsulation film, has also recently been described [103].

Some less commonly used, but nonetheless useful, analytical techniques include atomic force microscopy (AFM), electron spectroscopy for chemical analysis (ESCA, or XPS), and liquid chemical analysis techniques, such as ICPMS, ion chromatography (IC), and gas chromatography (GC).

Atomic force microscopy is used for measuring surface profiles and for quantifying variations such as surface roughness. The tip of a small probe at the end of a cantilever is brought very close (distance on the order of a nm) to a wafer and drawn laterally across the surface. Ideally, the tip should terminate with a single atom. The force between atoms on the surface and on the tip leads to bending of the cantilever. The degree of bending can be used to determine the force and to develop an image of the surface with resolution down to below nanometer scale [104]. The technique has been used for the study of effects of various wet chemical treatments, such as dilute and buffered hydrofluoric acid solutions [105], on roughening of silicon surfaces.

Electron spectroscopy for chemical analysis is a surface analysis technique based on the photoelectric effect. X-rays bombard a sample, resulting in photoemission of electrons, which are subsequently collected. Their spectrum indicates the elements present on the surface. The technique can be used for determining surface composition, analyzing surface contamination, and determining native oxide thickness.

Among liquid chemical analysis techniques, ICPMS is commonly used for determination of metal impurity levels. Solutions are introduced into an radio frequency (RF) plasma core, and ions produced in the plasma are analyzed by a mass spectrometer. Ion chromatography can be used for determining

assays of mixed acids and for determination of ionic contaminants in aqueous solutions, while GC can very sensitively determine chemical purity levels and give compositional information for solvent mixtures.

References

1. Kern, W., and D. A. Poutinen. "Cleaning Solution Based on Hydrogen Peroxide for Use in Silicon Semiconductor Technology." *RCA Rev.* 31 (1970): 187–205.
2. Weast, R. C. *CRC Handbook of Chemistry and Physics*, D-155–7. Boca Raton, FL: CRC Press, 1979.
3. Pourbaix, M., *Atlas of Electrochemical Equilibria in Aqueous Solutions*. Houston, TX: NACE, 1974.
4. Overton, T., and G. Rayner-Canham., *Descriptive Inorganic Chemistry*. 3rd ed. New York: W.H. Freeman, 2002.
5. Ohmi, T. "Cleaning Technology in Semiconductor Device Manufacturing IV." *ECS Proceeding*. Vol. PV95-20, 1. Pennington, NJ: The Electrochemical Society, 1996.
6. Tamilmani, S., W. Huang, S. Raghavan, and R. Small. "Potential-pH Diagrams of Interest to Chemical Mechanical Planarization of Copper." *J. Electrochem. Soc.* 149, no. 12 (2002): G638–42.
7. Vermeire, B., V. S. Pandit, H. G. Parks, S. Raghavan, R. Krishnaswami, and J. Jeon. "Hf or Zr High-k Fab Cross-Contamination Issues." *IEEE Trans. Semicond. Manuf.* 17, no. 4 (2004): 582.
8. Goh, F., C. Lim, V. K. T. Sih, I. Zainab, and S. Y. M. Chooi. "Occurrence of Arsenic-Based Defects and Techniques for Their Elimination." In *Proceedings, Seventh International Symposium on Ultra Clean Processing of Silicon Surfaces*. Brussels, Belgium, September 20–2, 2004.
9. Dean, J. A. *Lange's Handbook of Chemistry*, 5-7–12. New York: McGraw Hill, Table 5-12, 1979.
10. Rath, D. L. , Unpublished SC-1 modeling work.
11. Dean, J. A. *Lange's Handbook of Chemistry*, 5-49–53. New York: McGraw Hill, Table 5-14, 1979.
12. Sugiura, J. "Influence of Contaminants on Device Characteristics." In *Ultraclean Surface Processing of Silicon Wafers*, edited by T. Hattori, 29–41. Berlin: Springer, 1998.
13. *International Technology Roadmap for Semiconductors*. Austin, TX: Semiconductor Industry Association, 2003.
14. Zimon, A. D., *Adhesion of Dust and Powder*. New York: Plenum Press, 1969.
15. Hiemenz, P. C., *Principles of Colloid and Surface Chemistry*. New York: Marcel Dekker, 1986.
16. Donovan, R. P., and V. B. Menon. "Particle Deposition and Adhesion." In *Handbook of Semiconductor Wafer Cleaning Technology*, edited by W. Kern, Park Ridge, NJ: Noyes Publications, 1993.
17. Ouimet, G., D. L. Rath, S. L. Cohen, E. E. Fisch, and G. W. Gale. "Defect Reductiom and Cost Savings through Re-Inventing RCA Cleans." *Semiconductor Fabtech*, 5th ed., 305–13. London: ICG Publishing, 1996.
18. Schmidt, H. F., M. Meuris, P. W. Mertens, S. Verhaverbeke, M. M. Heyns, M. Kubota, and K. Dillenbeck. "Effect of Metallic Impurities on the Stability and Performance of Hydrogen-Peroxide Based Cleaning Solutions." In *Proceedings, Institute of Environmental Sciences 39th Annual Technical Meeting*, 238. Las Vegas, NV, May 2–7, 1993.
19. Schmidt, H. F., M. Meuris, P. W. Mertens, A. L. P. Rotondaro, M. M. Heyns, T. Q. Hurd, and Z. Hatcher. "H_2O_2 Decomposition and Its Impact on Silicon Surface Roughening and Gate Oxide Integrity." *Jpn. J. Appl. Phys.* 34 (1995): 727.
20. Storm, W., H. A. Gerber, G. F. Hohl, M. Naujok, and R. Schmolke. "Determination of SC-1 Etch Rates at Low Temperatures with Microscope Interferometry." In *Proceedings, Fourth International Symposium on Ultra Clean Processing of Silicon Surfaces*, 275. Ostend, Belgium, September 21–3, 1998.
21. Donovan, R. P., A. C. Clayton, D. J. Riley, R. G. Carbonell, and V. B. Menon. "Investigating Particle Deposition Mechanisms on Wafers Exposed to Aqueous Baths." *Microcontamination* 8, no. 2, (1990).

22. Riley, D. J., and R. G. Carbonell. "The Deposition of Liquid-Based Contaminants onto Silicon Surfaces." In *Proceedings of the Institute of Environmental Sciences 36th Annual Meeting*, 224–8. New Orleans, LA, 1900.
23. Gale, G. W. "Physical and Chemical Effects of High Frequency Ultrasound (Megasonics) on Liquid Based Cleaning of Si ⟨100⟩ Surfaces." PhD thesis, Clarkson University, 1995.
24. Ali, I. "Electrokinetic Characteristics of Particulate/Liquid Interfaces and Their Importance in Contamination from Semiconductor Process Liquids." PhD thesis, University of Arizona, 1990.
25. Riley, D. J., and R. G. Carbonell. "Mechanisms of Particle Deposition from Ultrapure Chemicals onto Semiconductor Wafers: Deposition from Bulk Liquid During Wafer Submersion." *J. Colloid Interface Sci.* 158 (1993): 259.
26. Cohen, S. L., W. A. Syverson, S. Basiliere, M. J. Fleming, B. Furman, C. Gow, K. Pope, R. Tsai, and M. Liehr. "Particle Removal Efficiency from Native Oxides Using Dilute SC-1 Megasonic Cleaning." In *Proceedings, Second International Symposium on Ultra Clean Processing of Silicon Surfaces*, 35. Bruges, Belgium, Sep. 19–21, 1994.
27. Schwartzman, S., A. Mayer, and W. Kern. "Megasonic Particle Removal from Solid-State Wafers." *RCA Rev.* 46 (1985): 81.
28. Kashkoush, I., A. Busnaina, F. Kern, and R. Kunesh. "Ultrasonic Cleaning of Surfaces: An Overview." In *Particles on Surfaces 3: Detection, Adhesion, and Removal*, edited by K. L. Mittal 217–37. New York: Plenum Press, 1991.
29. Gale, G. W., A. A. Busnaina, and I. Kashkoush. "Experimental Study of Ultrasonic and Megasonic Particle Removal." In *Proceedings, Precision Cleaning '94*, 232–53. Rosemont, IL, 1994.
30. Syverson, W., M. Fleming, and P. Schubring. "The Benefits of SC-1/SC-2 Megasonic Wafer Cleaning." In *Second International Symposium on Cleaning Technology in Semiconductor Manufacturing*, 10. Electrochemical Society Proceedings PV92-10, 1992.
31. Gale, G. W., and A. A. Busnaina. "Removal of Particulate Contaminants Using Ultrasonics and Megasonics: A Review." *Part. Sci. Technol.* 13 (1995): 197.
32. KeuhnD, T. H., D. B. Kittelson, Y. Wu, and R. Gouk. "Particle Removal from Semiconductor Wafers by Megasonic Cleaning." *J. Aerosol Sci.* 27, no. Suppl. 1 (1996): S427.
33. Mayer, A., S. Schwartzman. Megasonic cleaning system. U.S. Patent 3,893,769, Jul. 8, 1975.
34. Zhang, D. "Fundamental study of Megasonic cleaning." PhD thesis, University of Minnesota, 1993.
35. Gale, G. W., A. A. Busnaina, F. Dai, and I. Kashkoush. "How to Accomplish Effective Megasonic Particle Removal." *Semicond. Int.* 19, no. 9 (1996): 133.
36. Holsteyns, F., K. Lee, S. Graf, R. Palmans, G. Vereecke, and P. W. Mertens. "Megasonics: A Cavitation Driven Process." In *Proceedings, Seventh International Symposium on Ultra Clean Processing of Silicon Surfaces*. Brussels, Belgium, Sep. 20–2, 2004.
37. Lippert, A., P. Engesser, G. Farrell, J. Klitzke, M. Koffler, F. Kumnig, J. Leberzammer, et al. "Behavior of a Well-Designed Megasonic Cleaning System." In *Proceedings, Seventh International Symposium on Ultra Clean Processing of Silicon Surfaces*. Brussels, Belgium, Sep. 20–2, 2004.
38. Cohen, S. L., D. Rath, G. Lee, B. Furman, K. R. Pope, R. Tsai, W. Syverson, C. Gow, and M. Liehr. "Studies of the Relationship Between Megasonics, Surface Etching, and Particle Removal in SC-1 Solutions." In *Proceedings, Materials Research Society Symposium on Ultraclean Semiconductor Processing Technology and Surface Chemical Cleaning and Passivation*, 13–19. San Francisco, SA, Apr. 17–19, 1995.
39. Loper, S., and T. Wagener. "Minimizing Oxide Loss in Immersion SC-1 Processes." In *Proceedings, Eighth International Symposium on Cleaning Technology in Semiconductor Device Manufacturing*. Pennington, NJ: ECS Proceedings, Electrochemical Society, 2003.
40. Heyns, M. M., I. Cornelissen, S. De Gendt, R. Degraeve, D. M. Knotter, P. W. Mertens, S. Mertens, et al. "Advanced Cleaning and Ultra-Thin Oxide Technology." *SCP Global Technologies 5th International Symposium*. Boise, ID, Apr. 22–5, 1998.
41. Osaka, T., A. Okamoto, H. Kuniyasu, T. Hattori. "Single Wafer Spin Cleaning with Repetitive Use of Ozonated Water and Dilute HF." In *Proceedings, Seventh International Symposium on Cleaning Technology in Semiconductor Device Manufacturing*. Pennington, NJ: ECS Proceedings, Electrochemical Society, 2001.

42. Kanno, I., et al. In *Proceedings, Fifth International Symposium on Cleaning Technology in Semiconductor Device Manufacturing*. Vol. PV97-35, 54. ECS Proceedings, 1998.
43. Shah, R., J. J. Wu, C. Yu, C. H. Yang, D. Miura, N. Greco, and G. Gifford. "Cryogenic Aerosol Clean: A Novel Approach to Wafer Cleaning." In *Proceedings, Third International Symposium on Ultra Clean Processing of Silicon Surfaces*, 245. Antwerp, Belgium, Sep. 23–5, 1996.
44. Dean, J. A., *Lange's Handbook of Chemistry*. 12th ed. New York: McGraw-Hill, 1972 (Table 5-7).
45. Giguere, A., and Turrell. "The Nature of Hydrofluoric Acid, Aspectrographic Study of the Proton Transfer Complex $H_3O^+F^-$." *J. Am. Chem. Soc.* 102 (1980): 5473–77.
46. Henley, W., L. Jastrzebski, and N. Haddad. "Monitoring Iron Contamination in Silicon by Surface Photovoltage and Correlation to Gate Oxide Integrity." In *Proceedings, Materials Research Society Symposium*, 1993.
47. Meuris, M., M. Heyns, W. Kuper, S. Verhaverbeke, and A. Philipossian. "Correlation of Metal Impurity Content of ULSI Chemicals and Defect-Related Breakdown of Gate Oxides." *ULSI Science and Technology*. Vol. 9–11, 9. Pennington, NJ: ECS Proceedings, the Electrochemical Society, 1991.
48. Verhaverbeke, S., M. Meuris, P. W. Mertens, A. Kelleher, M. M. Heyns, R. F. DeKeersmaecker, M. Murell, and C. J. Sofield. "The Effect of Metallic Contamination on Void Formation, Dielectric Breakdown and Hole Trapping in Thermal SiO_2 Layers." *Cleaning Technology in Semiconductor Device Manufacturing*. Vol. 92-12, 187. Pennington, NJ: ECS Proceedings, the Electrochemical Society, 1992.
49. Kern, F. W., M. Itano, I. Kawanabe, M. Miyashita, R. Rosenberg, and T. Ohmi. "Metallic Contamination of Semiconductor from Processing Chemicals: The Unrecognized Potential." In *Proceedings, Advanced Wet Processing II*, 113. Tokyo, Japan: Ultra Clean Society, 1991.
50. Küpfer, W., and K. Maex. Presented at SEMICON Zürich, 1991.
51. Shimono, T., and M. Tsuji. *Extended Abstracts of the 179th ECS Meeting*. Washington, DC: ECS Proceedings, 1991.
52. Gräf, D., M. Brohl, and M. Bauer-Mayer. "Silicon Surface Treatments and Their Impact on Chemical Composition and Morphology." *Proceedings, Marterials Research Society Symposium* 315, (1993).
53. Rossiter, C., and I. I. Suni. "Atomic Force Microscopy of Au Deposition from Aqueous HF onto Si(111)." *Surf. Sci.* 430, no. 1–3 (1999): L553–7.
54. Watanabe, M., M. Hamano, and M. Harazono. *Mater. Sci. Eng.* B4 (1989): 401.
55. Park, J., and M. Pas. *J. Electrochem. Soc.* 142, no. 6 (1995): 2028.
56. Scriven, L. E., and C. V. Sternling. "The Marangoni Effects." *Nature* 187 (1960): 186.
57. Marra, J., and J. A. M. Huethorst. "Physical Principles of Marangoni Drying." *Langmuir* 7 (1991): 2748–55.
58. Wolke, K., B. Eitel, M. Schenkl, S. Ruemmelin, and R. Schild. "Marangoni Wafer Drying Avoids Disadvantages." *Solid State Technol.* 8 (1996): 87–90.
59. van Gelder, W., and V. E. Hauser. "The Etching of Silicon Nitride in Phosphoric Acid with Silicon Dioxide as a Mask." *J. Electrochem. Soc.* (1967): (Pennington, NJ).
60. Knotter, D. M., and T. J. J. Denteneer. "Etching Mechanism of Silicon Nitride in HF Based Solutions." *J. Electrochem. Soc.* 148, no. 3 (2001): F43–F46.
61. Morrison, R. T., and R. N. Boyd. *Organic Chemistry*. 3rd ed., 346. Boston, MA: Allyn Bacon, 1980.
62. Muti, C. J., and R. R. Matthews. "Chilled Ozone for Removing Photoresist Proves Practical." *Precision Cleaning* (1997): 11–5.
63. Gottschalk, G., U. Beuscher, S. Hardwick, M. Kobayashi, J. Schweckendiek, and M. Wikol. "Production of High Concentrations of Bubble-Free Dissolved Ozone in Water." In *Proceedings, Fourth International Symposium on Ultra Clean Processing of Silicon Surfaces*, 59. Ostend, Belgium, Sep. 21–3, 1998.
64. De Gendt, S., P. Snee, I. Cornelissen, M. Lux, R. Vos, P. W. Mertens, D. M. Knotter, M. M. Meuris, and M. Heyns. "A Novel Resist and Post-Etch Residue Removal Process Using Ozonated Chemistry." In *Proceedings, Fourth International Symposium on Ultra Clean Processing of Silicon Surfaces*, 165. Ostend, Belgium, Sep. 21–3, 1998.

65. Abe, H., H. Iwamoto, T. Toshima, T. Iino, and G. W. Gale. "Novel Photoresist Stripping Technology Using Ozone/Vaporized Water Mixture." *IEEE Trans. Semicond. Manuf.* 16, no. 3 (2003): 401–8.
66. Homma, Y., S. Kondo, N. Sakuma, K. Hinode, J. Noguchi, N. Ohashi, H. Yamaguchi, and N. Owada. *J. Electrochem. Soc.* 147, no. 3 (2000): 1193.
67. Beverina, A., H. Bernard, J. Palleau, J. Torres, and F. Tardif. *Electrochem. Solid-State Lett.* 3, no. 3 (2000): 156.
68. Kirkpatrick, B. K. "Surface Preparation Challenges with Cu/Low-k Damascene Structures." *ECS* Fall, (2001).
69. Archer, L., and T. Corteau. *Solid State Technol.* December, (2002).
70. Hartney, M. A., D. W. Hess, and D. S. Soane. *J. Vac. Sci. Technol.* B7 (1989): 1.
71. Wang, Y., S. W. Graham, L. Chan, and S.-T. Loong. *J. Electrochem. Soc.* 144 (1997): 1522.
72. Mannaert G., M. Van Cauwenberghe M.O. Schmidt J. Van Aelst D. Hendrickx M. Stucchi T. Conard S. Vanhaelemeersch W. Boullart "Resist Strip and Cu Diffusion Barrier Etch in Cu BEOL Integration Schemes in a Mattson Highlands™ Chambers." *Ultra Clean Processing Technology Symposium*, 16–18 September. Oostende, Belgium, 2002.
73. Savas, S., R. George, D. Gilbert, J. Cain, M. Herrick, A. Nagy, and K. Karuppana. *Micro* October/November, (2004).
74. Peters, L. "Low-k Drives New Stripping Solutions." *Semicond. Int.* 25, no. 12 (2002): 57–68.
75. Baklanov, M. R., Q. T. Le, E. Kesters, F. Iacopi, J. Van Aelst, H. Struyf, W. Boullart, and K. Vanhaelemeersch Maex. "Challenges of Clean/Strip Processing for Cu/Low-k Technology." In *Proceedings of the International Interconnect Technology Conference (IITC)*, 187–9. Piscataway, NJ: IEEE, 2004.
76. Mattox, D. M. *Handbook for Physical Vapor Deposition (PVD)*. William Andrew Publishing 1998, chap. 12.
77. Mungekar, H. P. and Y. S. Less. "Microelectronics and Nanometer Structures." *J. Vac. Sci. Technol. B* 24, no. 2 (2006): L11–15.
78. Lester, M. Semiconductor International, 1/12001, New Single-Wafer Processes Offer Alternative Backside Cleans.
79. Bowling, A., B. Kirkpatrick, T. Hurd, L. Losey, and P. Matz. "Future Challenges for Cleaning in Advanced Microelectronics." *UCPSS VI* 92 (2002): 1–6.
80. Wolke, K., B. Eitel, M. Schenkl, S. Rummelin, and R. Schild. "Marangoni Drying for Hydrophobic Wafers." *Semicond. Fabtech* (Edition 4, Section 7 Wafer Processing).
81. Zikanov, O., W. Boos, W. Wolke, and K. Thess. "A Model for Thermal Marangoni Drying." *J. Eng. Math.* 40 (2001): 249–67.
82. Mertens, P. W., G. Doumen, J. Lauerhaas, K. Kenis, W. Fyen, M. Meuris, S. Arnauts, K. Devriendt, R. Vos, and M. Heyns. "A High Performance Drying Method Enabling Clustered Single Wafer Wet Cleaning." In *Symposium on VLSI Technology Digest of Technical Papers*, 56–7. Leuven, Belgium: IMEC, 2000.
83. Lester, M. A. "Quick Drying Enables Single-Wafer Clearns." *Semicond. Int.* (2000).
84. Christenson, K., N. P. Lee, and T. J. Wagener. "Better HF Etch Uniformity with Single-Tank Approach." *Semicond. Int.* 25, no. 9 (2002): 46.
85. Thess, A., and W. Boos. "A Model for Marangoni Drying." *Phys. Fluids* 11 (1999): 3852.
86. Yalamanchili, M. R., and J. J. Rosato. "IPA Vapor Drying Technology to Meet Surface Preparation Challenges for Sub-0.18 μm Design Rules." *Future Fab Int.* 9, no. 20 (2000).
87. Woo-Gun, J., et al. "Evaluation of Reticle Cleaning Performance with Different Drying Methods for High-Grade Photomasks." *SPIE* 4562: 99–100, 21st Annual BACUS Symposium on Photomask Technology.
88. Goldfarb, D. L., J. J. de Pablo, P. F. Nealey, J. P. Simons, W. M. Moreau, and M. Angelopoulos. *J. Vac. Sci. Technol. B* 18, no. 6 (2000): 3313.
89. Namatsu, H. *J. Vac. Sci. Technol. B* 18, no. 6 (2000): 3308.
90. Wagner, M., J. DeYoung, S. Gross, Z. Hatcher, and C. Ma. "Cleaning Technology in Semiconductor Device Manufacturing VIII." Fall Meeting of the Electrochemical Society, Oct. 2003.
91. Levitin, G., S. Myneni, and D. Hess. *J. Electrochem. Soc.* 151, no. 6 (2004): 1.

92. Khare, J. B., W. Maly, and M. E. Thomas. "Extraction of Defect Size Distributions in an IC Layer Using Test Structure Data." *Semicond. Manuf., IEEE Trans.* (1994): 354–68 (Dept of Electr. and Computer Eng., Carnegie Mellon Univ, Pittsburgh, PA).
93. Stapper, C. H., and R. J. Rosner. "Integrated Circuit Yield Management and Yield Analysis: Development and Implementation." *IEEE Trans. Semicond. Manuf.* 8, no. 2 (1995): 95–102.
94. Lakhani, F., D. Dance, and R. Williams. "0.25-μm Integrated Circuit Yield Model Design and Validation." In *Proceedings of the IEEE International Symposium on Semiconductor Manufacturing*, E21–4. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 1997.
95. Ohring, M., *Reliability and Failure of Electronic Materials and Devices*. San Diego, CA: Academic Press, 1998.
96. Hu, C. K., and J. M. E. Harper. "Copper Interconnections and Reliability." *Mater. Chem. Phys.* 52 (1998): 5.
97. Yokogawa, S., N. Okada, Y. Kakuhara, and H. Takizawa. "Electromigration Performance of Multi-level Damascene Copper Interconnects." *Microelectron. Reliability* 41 (2001): 1409.
98. Hattori, T. "Detection and Identification of Particles on Silicon Surfaces." In *Particles on Surfaces 4: Detection, Adhesion, and Removal*, edited by K. L. Mittal, 208–17. New York: Marcel Dekker, 1995.
99. Hockett, R. S. "High Sensitivity Characterization of Contamination on Silicon Surfaces Using TXRF." In *Proceedings, Institute of Environmental Sciences 39th Annual Technical Meeting*, 432. Las Vegas, NV, May 2–7, 1993.
100. De Gendt, S., A. Huber, B. Onsia, S. Arnauts, K. Kenis, D. M. Knotter, P. W. Mertens, and M. M. Heyns. "Vapor Phase Decomposition—Droplet Collection: Can We Improve the Collection Efficiency for Copper Contamination?" In *Proceedings, Fourth International Symposium on Ultra Clean Processing of Silicon Surfaces*, 93. Ostend, Belgium, Sept. 21–3, 1998.
101. Adamson, A. W., *Physical Chemistry of Surfaces*. 5th ed. New York: Wiley, 1990.
102. Yabumoto, N. "Analysis and Evaluation of Molecules Adhered to Wafer Surfaces." In *Ultraclean Surface Processing of Silicon Wafers*, edited by T. Hattori, 179. Berlin: Springer, 1998.
103. Guan, J., G. W. Gale, and J. Bennett. "Effects of Wet Chemistry Pre-Gate Clean Strategies on the Organic Contamination of Gate Oxides." *SCP Global Technologies 6th International Symposium*. Boise, ID, May 10–2, 1999.
104. Hosaka, S. "Analysis of Microscopic Areas of Wafer Surfaces Using STM/AFM." In *Ultraclean Surface Processing of Silicon Wafers*, edited by T. Hattori, 223. Berlin: Springer, 1998.2
105. Bertagna, V., R. Erre, and M. Chemla. "Corrosion Rate of n- and p-Silicon Substrates in HF, HF+ HCl and HC+NH_4F Aqueous Solutions." *J. Electrochem. Soc.* 146, no. 1 (1999): 83.

6

Supercritical Carbon Dioxide in Semiconductor Cleaning

Mohammed J. Meziani
Pankaj Pathak
Ya-Ping Sun
Clemson University

6.1 Introduction **6**-1
6.2 Supercritical Fluids **6**-2
Characteristic Properties • Solvation Capabilities of Supercritical CO_2 in Semiconductor Cleaning
6.3 Supercritical CO_2 Cleaning Processes **6**-6
Photoresist Striping and Removal • Cleaning of Trace Metals and Particulates • Supercritical CO_2 in Etching • Supercritical CO_2 in Drying
6.4 Processing Equipment **6**-18
6.5 Conclusions and Perspectives **6**-20
References **6**-21

6.1 Introduction

The manufacturing of semiconductor devices relies on four basic operations: layering, patterning, doping, and heating.[1] In the layering operation, thin layers of a conductor, semiconductor, or nonconductor are added to the surface of a wafer. Patterning involves a series of steps, which ultimately result in the selective removal of the preciously deposited layers. Doping is a process that adds dopants to the wafer surface to change the conductivity of the semiconductor. The heating portion of the production process combines heating and cooling of the wafer to ensure good electrical conductivity. Each of these operations may be performed multiple times on a single wafer, and in each basic operation several procedures may be performed depending on the options selected for the type of circuit and its composition. Many cleaning procedures (at least one-third of all procedures) are typically involved in these operations.[1] It is well established that the device performance, reliability, and product yield are critically affected by the presence of chemical contaminants and particulate impurities on the device surface.[2] It is a generally accepted reality that more than half of the yield losses in integrated circuit fabrication are due to contamination from metallic, organic, and others.[3] Effective techniques for cleaning these devices before and after oxidation and patterning are now more important than ever before because of the extreme sensitivity of the semiconductor surface and the submicron sizes of the device features. The available techniques include cleaning with aqueous and organic solvents and the application of vapor-phase chemistries (both organic and inorganic), as well as the use of various

physical and thermal methods for contaminant removal.[2] Among widely discussed disadvantages for all of these methods are the consumption of large amount of high purity water, an excessive use of hazardous materials, such as acids, gases, and organic solvents, and the presence of multisteps.[4] In the current practice, hundreds and thousands of gallons of organic solvents and corrosive mixtures are used, and millions of gallons of waste water are produced in a standard chip manufacturing plant on an average operation day. In addition to these environmental concerns, there are also technical challenges for future operations.[4] To make faster and more powerful microelectronic devices, further reductions in the feature size and the incorporation of new materials are required.[4–6] The semiconductor industry is moving away from the traditional configuration of aluminum interconnect metal with silicon dioxide dielectric between the metal lines, and instead to use copper metal and low-κ (2.3 or lower) dielectric materials (often referred to as "dual damascene"). The more conductive copper reduces the resistance of the metal interconnect lines, while low-κ dielectrics reduce the parasitic capacitance between the metal lines. As the device size and features are reduced to nanoscale, it becomes extremely difficult for the use of liquid-based cleaning processes due to the high surface tension of liquids, especially with respect to reaching small and high aspect ratio trenches and the associated issues on a complete removal of residues. The drying process after liquid cleaning may also damage weak structures, such as microelectromechanical structures (MEMS) devices and porous low-κ dielectrics.

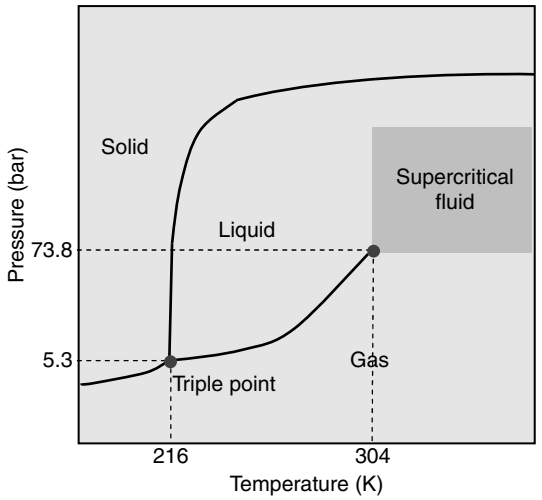
These environmental and technical challenges have prompted the semiconductor industry to adopt alternative and emerging technologies, including the use of carbon dioxide (CO_2) in semiconductor cleaning processes.[7–14] CO_2 is nontoxic, nonflammable, inert, and easy to recycle, thus reduced waste streams and significant cost savings.[15,16] Another technical advantage with the use of CO_2 for surface cleaning is that it is inert to inorganic materials.[17] It has been shown that supercritical CO_2 ($scCO_2$; purified and dry) has no corrosive action on stainless steel, iron, and copper.[18–23] All three states of CO_2 (liquid, gas, and supercritical) have found applications in precision cleaning. For example, liquid CO_2 is used for surface cleaning and degreasing.[7,8] CO_2 gas for dry ice particles (often called "snow" or "pellet") ejected from specialized nozzles;[7,8,24] and $scCO_2$ for chemical extraction cleaning.[7–14] These different states of CO_2 offer various options in the removal of organic contaminants, solid particulates, and deformable surface films or polymeric substrates, as produced in the photoresist development process. The cleaning with liquid CO_2 involves dipping and dissolving contaminated parts into a cleaning chamber, typically with agitation to increase the effectiveness of the cleaning process. The CO_2 snow relies primarily on the physical mechanisms for particle removal by dislodging the small, particulate matter, or dissolving organic oils into the liquid CO_2 from the instantaneous liquification of dry CO_2 on contact with the surface.[7,8,24] $ScCO_2$ is more flexible and advantageous than the other two states since it possesses properties intermediate between a liquid and a gas. In addition, as a result of its low viscosity and negligible surface tension, $scCO_2$ is capable of penetrating small pores and crevices and spreading out over surfaces easily like a gas, while still possessing the liquid-like solvent property to dissolve organic substances (for example, oil and grease). $ScCO_2$ cleaning also eliminates the need for drying, consequently increasing production rate and avoiding recontamination.

The focus of this review is obviously on $scCO_2$ cleaning. We first provide some background information on the characteristics of supercritical fluids and their related solvation properties, with an emphasis especially on $scCO_2$. We then highlight some novel applications of $scCO_2$ in the semiconductor cleaning process and recent advances in the emerging technology. We conclude with a brief summary of the challenges and perspectives in this interdisciplinary research field.

6.2 Supercritical Fluids

6.2.1 Characteristic Properties

A supercritical fluid may be defined as a gas or liquid at temperature and pressure above the critical point, where the fluid exists in a single phase.[25] The phenomenon is easily explained in reference to the phase diagram for pure CO_2 (Figure 6.1). It maps the regions of pressure and temperature over which the



	Liquid	Vapor	Critical point	Supercritical fluid
Density (g/cm^2)	1.0	~10^{-3}	0.5	0.3–0.7
Diffusivity (cm^2/s)	<10^{-5}	~10^{-1}	~10^{-3}	10^{-2}–10^{-5}
Viscosity (g/cm s)	~10^{-2}	~10^{-6}	~10^{-2}	10^{-3}–10^{-6}

FIGURE 6.1 The phase diagram of carbon dioxide (CO_2) and a table of physical properties.

various phases are thermodynamically stable.[26] At the critical point, the density of the gas phase becomes equal to that of the liquid phase and the interface between gas and liquid disappears. A comparison of typical values for density (ρ), viscosity (η), and diffusivity (D) of gases, liquids, and supercritical fluids is also presented in Figure 6.1. The physical properties of a supercritical fluid are intermediate between those of the liquid and gas phases, with their transport properties (diffusivity and viscosity) similar to those of gases but solvating powers similar to those of liquids. Among the most important properties of a supercritical fluid are the low and tunable densities, which can be easily varied from gas- to liquid-like via a simple change in pressure at constant temperature, and the unusual solvation effects at densities near the critical density (often discussed in terms of solute–solvent clustering and solute–solute clustering). Commonly used supercritical solvents include CO_2, ethylene, ethane, fluoroform, and ammonia.

A supercritical fluid may be considered as being macroscopically homogeneous but microscopically inhomogeneous, consisting of clusters of solvent molecules and free volumes.[27,28] With the macroscopic homogeneity, extremely wide variations in solvent properties may be achieved in a single supercritical fluid via changes in the fluid density. In fact, many investigations have focused on the dependence of solvation properties on density in supercritical fluids. For example, spectroscopic techniques, coupled with well-characterized molecular probes, have been employed to examine the local polarities at different densities in a supercritical fluid.[29–31] Results from these investigations suggest that the solvent strength of a supercritical fluid increases with increasing fluid density in a nonlinear fashion, deviating significantly from the prediction based on the classical dielectric-continuum theory.[32–34] Interestingly, however, the pattern for the density dependence of solvation properties (determined using molecular probes based on drastically different mechanisms) is nearly universal among supercritical fluids in different categories (from nonpolar to polar and from ambient to high temperature). Based on these results, Sun and co-workers[32] proposed a three-density-region solvation model for solute–solvent interactions in supercritical fluids (Figure 6.2). Experimentally, the solvent effects are strong (increasing significantly with density) in the gas-like region (approximately, reduced density $\rho_r < 0.5$), nearly plateau-like in the near-critical region (approximately, $0.5 < \rho_r < 1.5$), and moderately density-dependent in the liquid-like region ($\rho_r > 1.5$).[32–34] According to this model, the density dependence of solvation in supercritical fluid solutions is governed by the intrinsic properties of the neat fluid over the three-density regions. The behavior in the gas-like region at low densities is probably strongly influenced by short-range interactions in the inner solvation shell of the probe molecule. The strong density dependence of the spectroscopic and other responses is probably associated with a process of saturation of the inner solvation shell. Before

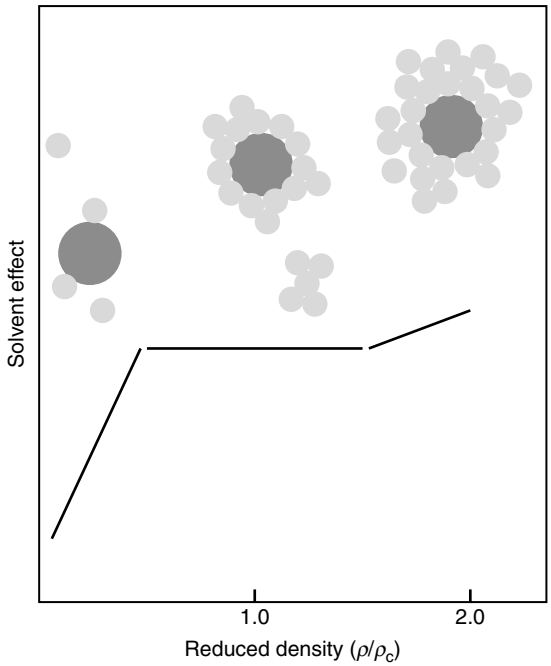


FIGURE 6.2 A cartoon illustration of the three-density-region solvation model. (From Sun, Y.-P., Bunker, C.E., and Hamilton, N.B., *Chem. Phys. Lett.*, 210, 111, 1993.)

saturation of the inner shell, the consequence of increasing the fluid density is microscopically the addition of solvent molecules to the inner solvation shell of the probe, which produces large incremental effects (Figure 6.2). In the near-critical region where the responses are nearly independent of density changes, the microscopic solvation environment of the solute probe undergoes only minor changes. Such behavior is probably due to the microscopic inhomogeneity of the near-critical fluid—a property that all supercritical fluids share. Despite the dynamics, the fluid in the near-critical region can on average be viewed as consisting of solvent clusters and free volumes that possess liquid- and gas-like properties, respectively. Changes in bulk density through compression primarily correspond to decreases in the free volumes, with solute–solvent interactions in the solvent clusters being largely unaffected. At the boundary of this region, the free volumes become less significant (consumed), and further increases in bulk density in the liquid-like region alter the microscopic solvation environment of the probe in a manner similar to that in normal liquid solvents, as predicted by the classical dielectric-continuum theory.

In addition to the solute–solvent interactions, the effect of solvent local-density augmentation on solute–solute interactions in a supercritical fluid solution has been the subject of extensive investigations,[35] with the focus being on whether the supercritical solvent environment facilitates solute–solute clustering, which may be loosely defined as the local solute concentration being higher than the bulk solute concentration. An important consequence of solute–solute clustering is the enhancement of bimolecular reactions in supercritical fluid solutions, which provides potentially significant opportunities for manipulating chemical reactions and processes under supercritical fluid conditions. The investigations employed well-established probes that are sensitive to bimolecular processes. The results seem to suggest that the solute–solute clustering is system-dependent, which makes it difficult to confirm experimentally the existence of local concentration augmentation or solute–solute clustering in an unambiguous fashion. Thus, the effect of supercritical solvent environment on solute–solute interactions remains a somewhat controversial topic.

6.2.2 Solvation Capabilities of Supercritical CO_2 in Semiconductor Cleaning

The characteristic properties of supercritical fluids discussed above and their ability to solvate or precipitate solutes selectively have made them uniquely applicable in chemical reactions and materials processing. Most of the research and development in these areas have been focused on $scCO_2$ because of the cost consideration and established knowledge base. Examples for applications based on the advantageous properties include the ability to tune selectively chemical reactions or processes,[36–41] the enhancement of reaction rates due to the low viscosities or high diffusivities in the fluids,[40,41] the production of powders and fibers via rapidly expanding the supercritical fluid solutions,[28,35] separating and extracting chemicals,[42–46] synthesizing polymers and pharmaceuticals,[47–49] degreasing machined parts, and dry cleaning.[7] Supercritical fluids have also attracted gradually increasing attention in the semiconductor industry, especially for addressing many of the challenges in microelectronics cleaning.[7–14] As the target dimensions in chip manufacturing continue to decrease, only solvent in the supercritical state is able to reach into the vias and trenches of integrated circuits. The gas-like diffusivity and low surface tension combined with liquid-like densities would significantly enhance the wetting capability and cleaning effectiveness for complex substrates with intricate geometries that trap contaminants. Another important advantage is to use the temperature and pressure dependence of solubility in supercritical fluid extraction to capture and condense contaminants. Contaminants found on semiconductor parts can be either organic or inorganic and in the form of films or particles on the surface. The cleaning procedure must be well designed in order to be able to handle both the chemical type (organic or inorganic) and the physical form (film, particulate, or surface incorporated) of the contamination.

$ScCO_2$ is obviously an ideal choice in the cleaning process. It has a near-ambient critical temperature ($\sim$31°C) and a relatively low critical pressure (73.8 bar), and it is nontoxic, nonflammable, and abundant. $ScCO_2$ behaves as nonpolar organic solvent and is therefore good for dissolving nonpolar organic compounds, such as greases, oils, lubricants, and fingerprints. For less or insoluble contaminants such as polar compounds, new formulations have been developed by adding modifiers to $scCO_2$, which include co-solvents, surfactants, chelating agents, and/or chemical reactants.[50] Polar co-solvents are often added to $scCO_2$ to enhance the solubility of organic residues.[51,52] This has been useful in cleaning applications, such as the removal of photoresists and post-etch/ash residues.

Formulated surfactants are also added in small quantity to $scCO_2$ to greatly enhance the solvation power in the solution to support ionic dissociation and to enable the dissolution of a wide range of polar species, such as those found in post-etch and post-ash residue removal and metal cleaning. The surfactants in $scCO_2$ stabilize microemulsion systems that contain discrete nano- or microscale polar domains of pure water or aqueous solutions dispersed in the continuous CO_2 phase. As illustrated in Figure 6.3, the

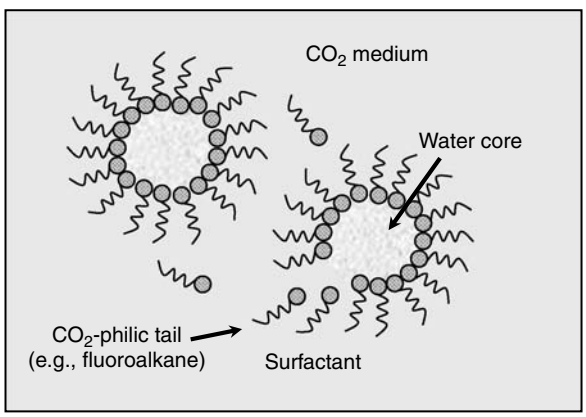


FIGURE 6.3 A schematic illustration on the structure of water-in-CO_2 microemulsion. The aqueous core is surrounded by surfactant molecules consisting of polar heads and CO_2-philic tails.

surfactant molecules form organized molecular aggregates and preferentially occupy the interface between CO_2 and water phases. The ability of the microemulsion to dissolve polar solutes depends solely on the characteristics of the microemulsion droplets. These polar droplets are able to simultaneously deliver active chemicals to the substrate surface as well as to solubilize and transport unwanted materials away from the surface. The use of active chemicals, such as acids, bases, chelators, etchant, and oxidants with water-in-CO_2 microemulsion systems is highly efficient for rapid and complete removal of residues. There has been considerable progress in the design and synthesis of CO_2-soluble surfactants.[53–62] The successful ones typically contain a CO_2-philic segment, such as a fluoroether-, fluoroacrylate-, or silicone-based compound and a CO_2-phobic segment of hydrophilic or lipophilic molecules.[53–64] Some popular examples are perfluoropolyether (PFPE), fluorinated AOT (sodium bis(2-ethylhexyl) sulfosuccinate, also called Aerosol-OT) analogs, and fluorinated phosphates. A PFPE tail with an average molecular weight of 2500 g/mol led to water-in-CO_2 microemulsions with the highest water content.[63] The surfactant head group was found to have a significant effect. Nonionic carboxylic acids did not form micelles in dense CO_2 but cationic species, such as PFPE-C(O)-NH-CH_2-$N^+(CH_3)_3CH_3COO^-$ led to the formation of water-in-CO_2 emulsions with $W_o \sim 28$ and droplet radii in the range 1.6–3.6 nm.[64] Recently, water-in-CO_2 microemulsions with large W_os (up to 28) have also been formed with nonionic methylated branched hydrocarbon surfactants, poly(ethylene glycol) 2,6,8-trimethyl-4-nonyl ethers, at moderate conditions (35°C–65°C and above 240 bar).[57] This is unprecedented for the use of a low molecular weight hydrocarbon surfactant in CO_2, with the result being comparable to those of fluoroether surfactants.

6.3 Supercritical CO_2 Cleaning Processes

New technology solutions and processes in wafer cleaning have become essential to meet the International Technology Roadmap for Semiconductors requirement for reduced surface contamination.[5] The presence of contaminants degrades the device quality and reliability and affects the overall device yield. These technological challenges have already been summarized and discussed in the literature.[65,66] The continuous downward scaling of devices to below 100 nm and the introduction of new materials in advanced devices now require the nonetching and damage-free techniques for precise interface control. Nonliquid wafer cleaning techniques with the use of $scCO_2$ are still under development, but already show great promise to become mainstream cleaning processes. $ScCO_2$, with its tunable density, low viscosity, and negligible surface tension, will likely provide the semiconductor industry with an integrated solution for the post-etch residue cleaning and the drying of porous low-κ materials. The technique is also flexible, with the use of precise formulations to target-specific applications, including photoresist development and removal, particle removal, etching, drying, etc.

6.3.1 Photoresist Striping and Removal

Photoresist is defined as a light sensitive chemical material that undergoes an as-designed change in chemical properties after light exposure.[1] It is used as sacrificial layer in the process of photolithography to create the desired patterned coating on wafer surface. The patterning is made through several fabrication steps by applying and selectively removing the photoresist, and a new layer is applied to the wafer in each step (Figure 6.4). The process begins with the covering of the whole substrate surface with a thin layer of photoresist material, usually via spin-coating, and then the coating is selectively irradiated with light (usually UV) through a negative or mask that is designed to allow light exposure for only the desired areas. Depending on the system, it is possible to solubilize and wash away either the exposed or unexposed regions selectively using the appropriate fluid, called the developer. When the exposed region is removed by the developer, the process is called positive tone, and when the developer leaves the irradiated region behind, the process is called negative tone. The exact composition and makeup of the photoresist vary widely.[67–69] The earliest photoresists were based on bis-azide photocross-linking. After the polymer is cross-linked, it becomes less soluble, ready for negative development. Another interesting

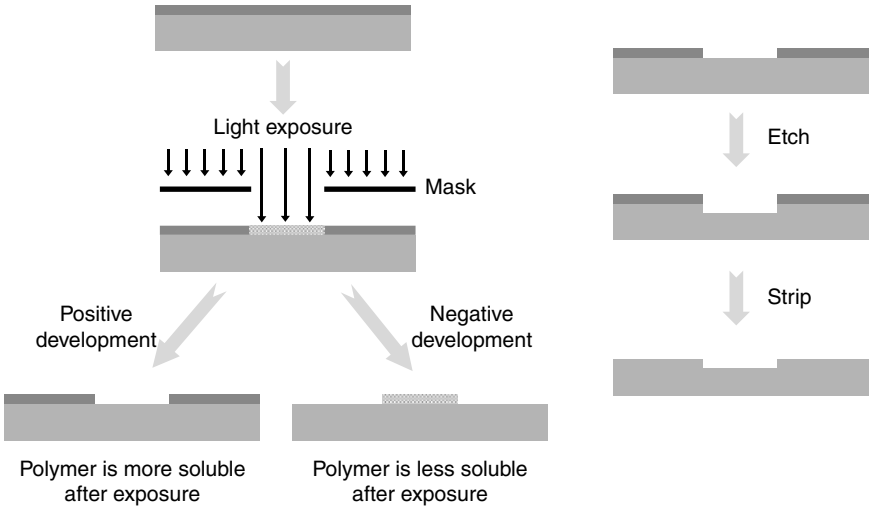


FIGURE 6.4 A scheme of typical lithographic processing. (From www.chem.rochester.edu/~chem421/polymod2.htm.)

negative-tone photoresist is poly(vinyl cinnamate), where the cinnamate groups undergo [2+2] cycloaddition when irradiated, leading to a cross-linked polymer. A popular and effective positive tone photoresist is Novolaks, containing Novolak polymer with a small amount of dissolved diazonaphthaquinone. When irradiated, the diazonaphthaquinone undergoes the photochemical Wolf rearrangement, which eventually produces a carboxylic acid easily soluble in base, leaving behind the insoluble unirradiated regions. Intricate features can be produced with the use of this photoresist.

The removal of photoresists is currently considered as a major application problem for the semiconductor industry. Most of the existing techniques for photoresist removal are based on wet-chemical treatments, plasma-induced dry ashing, or their combination. In a wet treatment procedure, most common wash systems are aqueous acidic or alkaline solutions and organic solvents, depending on the type of resist. For example, in traditional way, the wafers are dipped in a harsh chemical bath of sulfuric acid and hydrogen peroxide or sulfuric acid with ozone, followed by heating or ultrasonication. While these solutions are effective, they are environmentally undesired. The operating costs for these processes are also enormous, with expenses including the consuming of a large amount of ultra-pure water for rinsing, the need for special handling in the storage, and treating and disposing of post-process solvents and wastewater. The wet-chemical stripping processes are approaching their limits as well, as the chip architecture gets smaller.[70] For the dry stripping method, in which a dry plasma process oxidizes the photoresist into gases for removal via ventilation of the plasma chamber,[1] the advantage is the elimination of solvent use and the need for chemical hoods as in the wet stripping. However, there are also problems associated with this process, including incomplete resist removal and undesired byproducts, such as oxidizing gas and dry-etch residue. Another more significant downfall to this process is the undesirable effect on the electrical properties and integrity of the wafer because of the use of radiation to generate the plasma field.[1]

The recent introduction of $scCO_2$ as an alternative green method for photoresist removal promises great benefits to the industry and to the protection of environment. Primary reasons behind the use of $scCO_2$ include the potential to eliminate concerns over residual solvent, high-temperature plasma ashing, and pore collapse due to capillary forces, and the potential to allow photoresist removal in one single step at low temperature. An important issue in the use of $scCO_2$ stripping process is to ensure its compatibility with the substrate materials. Significant efforts have been made on developing new

lithographic systems based on polymeric resist materials of enhanced solubility in liquid and $scCO_2$.[71–83] A variety of fluoro- and silicon-containing polymers have been tested as positive- and negative-tone resists (Figure 6.5). An initial investigation was to use $scCO_2$ as solvent to extract unreacted oligosiloxane compounds from an organic polymer matrix.[71] Another early attempt to construct photoresists was based on the conversion of poly(silanes) into poly(siloxanes) by the photo-induced insertion of oxygen. These polymers, changing from being insoluble to soluble in $scCO_2$, were successfully imaged as positive tone resists.[71–73] The results demonstrated good contrast and excellent sensitivity of silicon-containing conversion of poly(silanes) into poly(siloxanes) by the photo-induced insertion of oxygen. For negative-tone CO_2 developable photoresists, Ober et al.[74] designed imageable copolymers from combinations of *t*-butyl methacrylate (*t*-BMA) with either 3-methacryloxypropylpentamethyl-disiloxane (SiMa) or pentafluoropropyl methacrylate (PFM). While poly(*t*-BMA) homopolymer is insoluble in $scCO_2$, copolymers of *t*-BMA with PFM and SiMA are soluble in $scCO_2$. The same group designed CO_2-soluble fluoromethacrylate block copolymers to be used as high-resolution, chemically amplified photoresists capable of imaging with 193-nm radiation.[79] It allowed the E-beam patterning (150–200 nm) and the development of an "all dry" lithography process. In a similar approach, other photoresists were synthesized based on random copolymers of 1,1-dihydroper-fluorooctylmethacrylate and 2-tetrahyrdopyranyl methacrylate.[80] These resins, along with specially designed CO_2-soluble photoacid generators (PAGs), were utilized to demonstrate the potential of a new "dry" lithographic process. The change in chemical properties of these photoresists is often promoted through the presence of a PAG that catalyzes the desired transformation. PAG is incorporated to form acid when exposed to radiation at a selected wavelength. The acid reacts with an acid-labile group, such as the tetrahydropyranyl or *t*-butyl group to generate an unprotected group. In carefully desired polar PAG-polymer combinations, the PAG may selectively partition to the photoacid cleavable block to make the deprotection more efficient. The CO_2-based photoresist removal has been shown to be compatible with various process conditions, such as the use of co-solvent with $scCO_2$ and the sharp pressure cycles applied to porous ultra low-κ and copper.[84,85]

The removal of other difficult or hardened photoresists relies on the dissolution of $scCO_2$ into the polymer. In this approach, the fluid attacks the bonds at the wafer–polymer and polymer–polymer interfaces and in essence floats the photoresist off the surface. It is well known that the interaction of CO_2 with polymeric substrates at high pressure can result in several effects, such as swelling, crystallization, and selective extraction.[86–88] The dissolution of CO_2 in polymeric matrix swells the polymer, especially if the polymer is kept slightly above its glass transition (T_g) but below its melt temperature (T_m). The swelling reduces both polymer–polymer and polymer–substrate interactions, promoting debonding. The combination of $scCO_2$ with a minimal amount of co-solvent along with pressure fluctuation can further induce swelling and debonding of the film. Under these conditions, $scCO_2$ acts as "carrier" to deliver the co-solvent to areas where cleaning is needed, and also take it back out. Co-solvents selected for use in low concentrations include propylene carbonate, DMSO, acetyl acetone, and acetic acid, possibly mixed with amines.[51,52] Several research groups have pursued this approach, including for example, You[89] and Biberger at Supercritical Systems, Inc.[90] Other original studies were conducted at Los Alamos National Laboratory (LANL) and industrial partners for the development of a treatment system called $ScCO_2$ Resist Remover (SCORR).[91–94] $ScCO_2$ was mixed with a small amount of propylene carbonate (typically 5% or less) in a pulsed flow system for the successful removal of hard-baked photoresist from aluminum-metallized Si wafers without causing any damage to the thin aluminum metallization patterns. Propylene carbonate was initially discovered as an alternative photoresist remover to replace the highly toxic methylene chloride and methyl chloroform,[95] despite the slightly less efficiency. It is a solvent of low toxicity and low vapor pressure, and it is nonflammable and completely miscible with high-pressure CO_2.[96] In an earlier study of using propylene carbonate with CO_2 under supercritical conditions, it was found that the solubility was higher in the mixtures than in either propylene carbonate or CO_2 alone.[97]

A schematic illustration of the SCORR system and related experimental process flow is provided in Figure 6.6.[11,90–93] The SCORR solvent is supplied by a cylinder containing the pre-mixed compressed

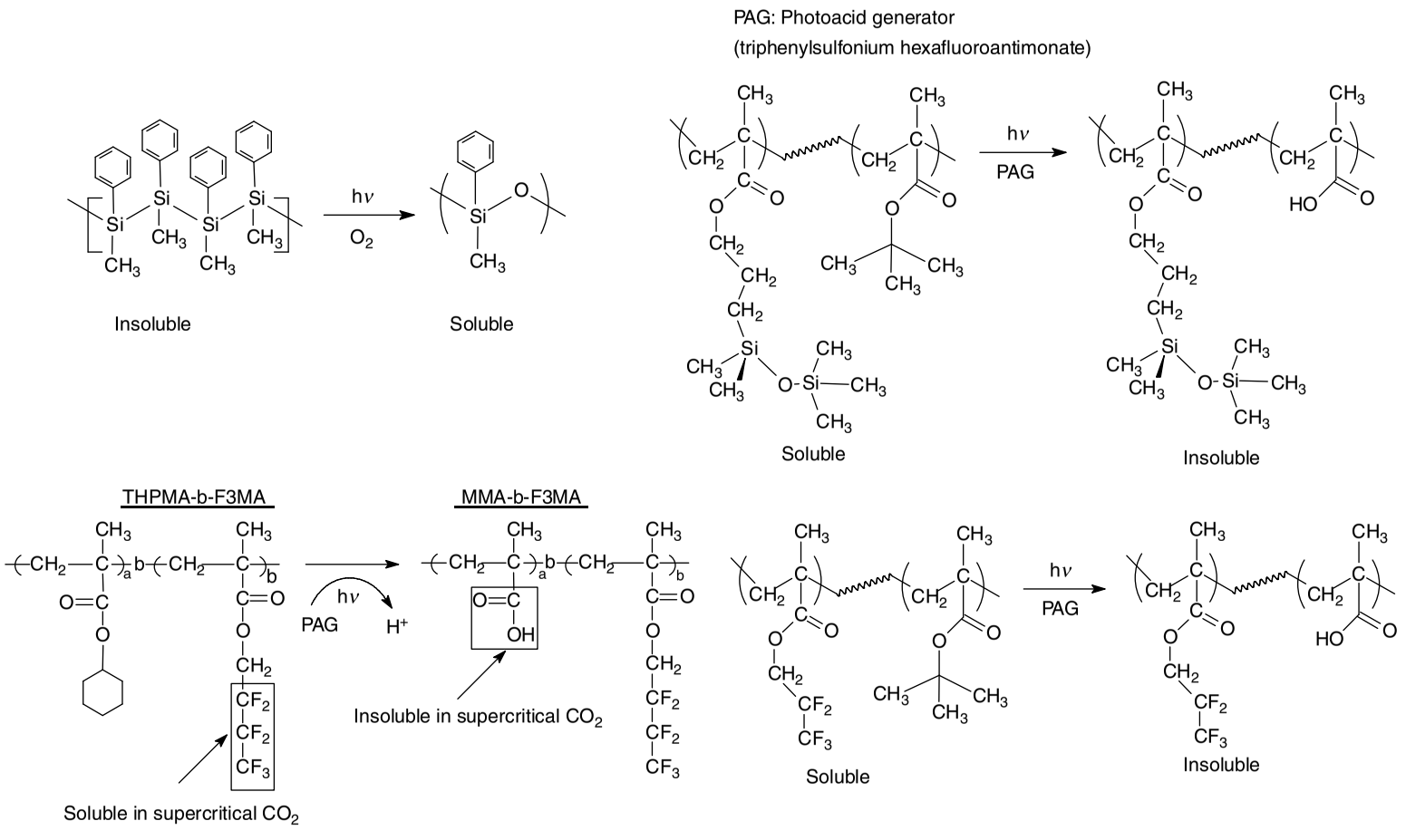


FIGURE 6.5 Representative CO_2-soluble photoresist polymers. (From Ober, C.K., Gabor, A.H., Gallagher-Wetmore, P., and Allen, R.D., *Adv. Mater.*, 9, 1039, 1997; Sundararajan, N., Yang, S., Ogino, K., Valiyaveettil, S., Wang, J.-G., Zhou, X., Ober, C.K., Obendorf, S.K., and Allen, R.D., *Chem. Mater.*, 1, 41, 2000.)

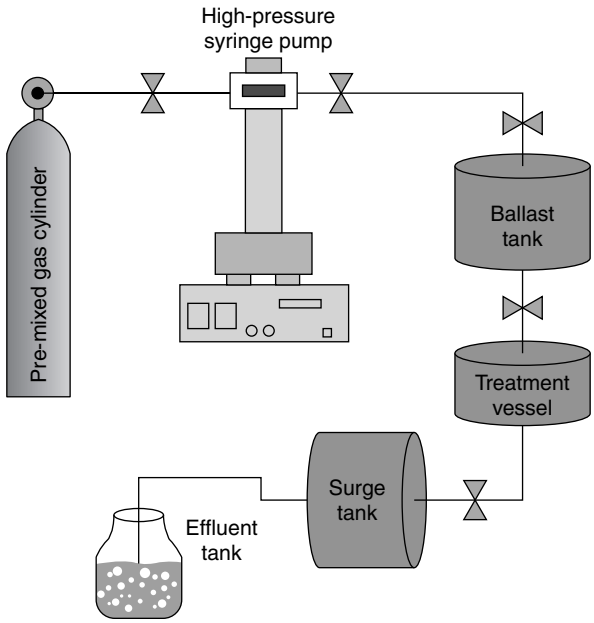


FIGURE 6.6 A setup for photoresist stripping. (From King, J.W. and Williams, L.L., *Curr. Opin. Solid State Mater. Sci.*, 7, 413, 2003.)

through a high-pressure syringe pump. A sample of scribed wafer (half-inch square) is centrally mounted inside the treatment vessel. The initial conditions for the ballast tank and treatment vessel are 11 and 7.6 MPa, respectively, both at 50°C. A valve leading from the ballast tank to the treatment vessel is opened, allowing the solvent mixture to flow through a nozzle into the vessel and onto the sample surface. This flow continues for several seconds, until the equilibrium in pressure (about 1400 psi) is established in the tank and vessel. Then the treatment vessel is depressurized back to 1100 psi by opening a needle valve from the treatment vessel to the surge tank, and the ballast tank is repressurized to 1600 psi. This pressurization/depressurization cycle is repeated several times, along with applying the SCORR solvent during the pulsation cycle. Finally, the supply of propylene carbonate is shut off, and a pure CO_2 "rinse and dry" step is applied to the treated sample for cleaning. The SCORR process has been demonstrated for wafers of GaAs, GaP, Si with Al, Ti–Pt, Ti–W, and In–Sn oxide metallizations covered with positive photoresist (AZ-4330).[94] Presented in Figure 6.7 are Scanning electron microscopy (SEM) images showing the effectiveness of the SCORR process in the cleaning of fluorinated photoresists on semiconductor surfaces. The results were obtained by employing $scCO_2$ at 20 MPa and 85°C, with either propylene or butylene carbonate as a co-solvent in concentrations of 1.0–3.8 vol.%.[11] These experimental conditions were also found effective for delaminating poly(methylmethacrylate) and Novalac resin from silica and Al/C surfaces. It was also demonstrated that the overall treatment time for photoresist softening could be reduced significantly with an increase in the treatment temperature.

The overall effectiveness of CO_2 with co-solvent in the pulsed mode for the cleaning and stripping of photoresist, even in the "trench" profiles, has also been confirmed by results from the surface characterization with the use of Fourier transform infrared (FTIR), Auger electron spectroscopy, and nuclear reaction analysis (NRA).[91] The mechanism by which the above process occurs is still not completely clear, but likely appears to be a combination of three physico-chemical effects: swelling of the polymer film, reduction in the glass transition temperature of the polymer, and possible degradation of the polymeric matrix due to the effects of pressure and the presence of a reactive ester group in propylene carbonate.[11]

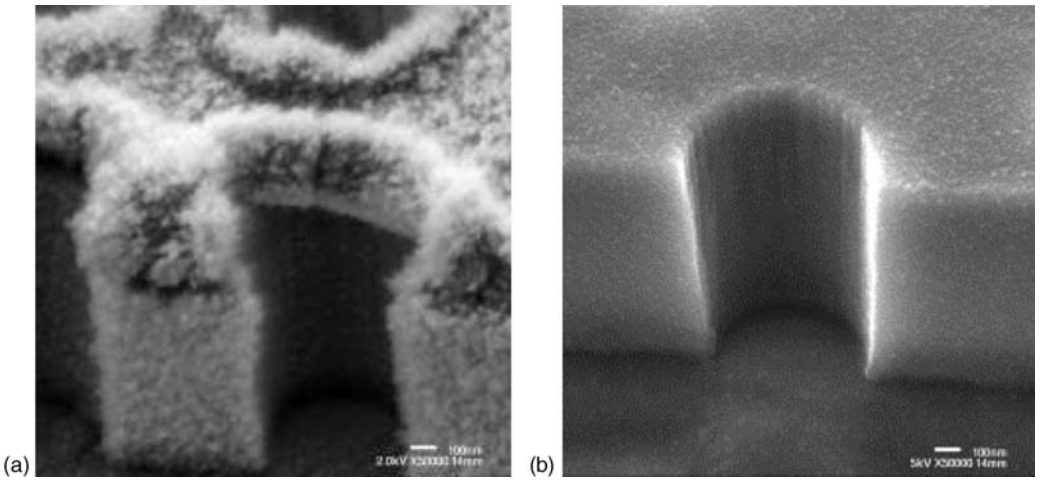


FIGURE 6.7 SEM images on the cross section of a wafer via (a) before and (b) after a complete removal of photoresist film using $scCO_2$ in the $ScCO_2$ Resist Remover (SCORR) process. (From King, J.W. and Williams, L.L., *Curr. Opin. Solid State Mater. Sci.*, 7, 413, 2003.)

6.3.2 Cleaning of Trace Metals and Particulates

Particle contaminants on semiconductor wafer surfaces can be either charged (ionic) or uncharged.[2] The contaminants can be brought along with the organic residues or deposited from equipment, manufacturing processes, factory operators, gas piping, etc. Ionic species comprise cations and anions, mostly from inorganic compounds that may be physically adsorbed or chemically bonded (chemisorbed), such as ions of sodium, fluorine, and chlorine. Ionic contaminants cause undesired effects in semiconductor devices. In the high-temperature processing or in the presence of an electric field, they may diffuse and spread on the surface of the semiconductor structure, leading to electrical defects, device degradation, and yield losses. Other undesired phenomena, such as twinning dislocations, stacking faults, and other crystal defects, can occur at sufficiently high ion concentrations. Uncharged species comprise metals, such as gold and copper, that are chemically or electrochemically plated out on the semiconductor surface from acid etchant solutions, or silicon particles or metal debris from equipment. These impurities if not removed may diffuse into the silicon substrate during later heat treatments and cause various device degradation and reliability problems, such as uncontrolled drifts in the semiconductor surface potential and excessive leakage currents. Metal contaminants in or on semiconductor wafers can also lead to structural defects in vapor-grown epitaxial layers and degrade the breakdown voltage of gate oxides. The uncharged contaminants are difficult to remove because they are less soluble than ions and generally require oxidation to make them soluble. A more detailed discussion on the adhesion forces that hold contaminants on surfaces is available in the literature.[98] It is generally assumed that the attractive forces between particle and surface are van der Waals, electrostatic, and magnetic in nature. The problem of particle contaminants becomes exacerbated as the particle sizes decrease because of the extremely strong adhesion forces.

Conventional approaches used to remove such contaminants from semiconductor wafers are liquid- and gas-phase cleaning methods. Different process combinations and sequences are employed for specific applications.[2] Liquid cleaning methods for semiconductor wafers are based on the application of an organic solvent or inorganic acid containing a small amount of surfactants, corrosion inhibitors, and/or complexation agents. The mechanism of liquid cleaning is primarily physical dissolution and/or chemical reaction dissolution. A number of vapor-phase cleaning processes for the removal of various groups of contaminant types have also been used.[2] Physisorbed and chemisorbed ions and deposited elemental

metals require chemical processes to remove them from the semiconductor or oxide surfaces. Physical enhancement techniques are frequently used to decrease the thermal energy requirement by supplying electromagnetic radiation or energy from plasma environments. The key requirement for removing chemical impurities is the formation of volatilizable species by reaction at low temperature, followed by their elimination at low pressure and an elevated temperature. In addition to the high cost and the generation of a large volume of waste, which must be disposed safely, these conventional approaches also face challenges with dewetting of nonpolar surfaces, highly porous structures, damage by plasmas, and change in k due to absorption of chemicals.

$ScCO_2$-based formulations can be used to address a number of these challenges, including easier handling and cleaning of residues. Since CO_2 is a linear molecule with no dipole moment, $scCO_2$ alone is a poor solvent for dissolving polar compounds and ionic species. The compounds that are gaseous at $scCO_2$ temperature and pressure, such as SF_6, BF_3, and C_2F_6, are readily soluble, but for liquid- and solid-state compounds the situation is somewhat more complex. Neat dense CO_2 is only capable of dissolving nonpolar and slightly polar molecular species and some inorganic compounds of low molecular weight. For example, $TiCl_4$ and $SnCl_4$ have been shown to be highly soluble in $scCO_2$.[99,100] A significant development in dispersing and transporting highly polar compounds and ionic residues has been the use of water-in-CO_2 microemulsion systems. Johnston and co-workers investigated these systems in the removal of post-etch residues from nanoscale vias and trenches in low-κ-patterned porous methylsilsesquioxane (pMSQ) interlayer dielectrics.[101] They found that the removal was effective with the use of a stable microemulsion containing CO_2, water, and a hydrocarbon surfactant poly(ethylene glycol) 2,6,8-trimethyl-4-nonyl ether. In contrast, the residues could not be removed using only water, CO_2 (with or without co-solvent), or dry CO_2 with surfactant. No collapsing or voiding of the pores was observed according to spectroscopic ellipsometry and SEM results upon pressurization and depressurization with the use of CO_2. A mechanistical explanation was that the particles were removed both as a suspension and in the molecularly dissolved state in micro- and macroemulsions.[101]

The switching to the new generation of copper and low-κ technology in dual damascene processing presents additional challenges, demanding special attention on backside Cu contamination (known as breakthrough and post-etch and post-ash residues). As an example, shown in Figure 6.8 is a portion of the process flow to create via first dual damascene structure.[102] After etching the low-κ film to define the trench, ashing to remove photoresist, and then sputtering to remove oxides on the copper metal at the bottom of the via, residues are formed on the sidewalls and within the pore of the low-κ film.[102] The residues containing oxidized photoresist fragments and copper metals must be removed so that the subsequently deposited barrier metal layer adheres to the low-κ film. Copper ions are often present on the surfaces of dielectric sidewalls and on the top of Cu(0) lines after the bottom barrier is etched away. These residues in interconnect processing can directly affect yields, and therefore their removal is

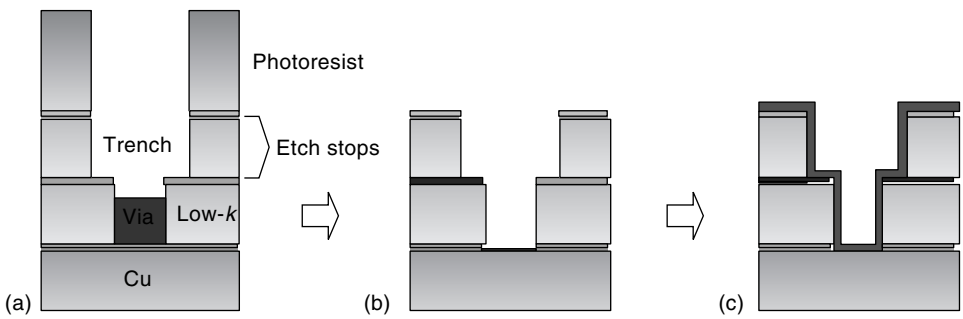


FIGURE 6.8 Portion of a via first dual damascene flow: (a) after etching trench; (b) after ashing photoresist, opening via, sputtering to clean copper at the bottom of the via and cleaning low-κ film; and (c) after depositing a barrier metal and copper seed layer. (From Muscat, A.J., *Business Briefing: Global Semicond. Manuf. Technol.*, 2003.)

necessary. Researchers at Micell Integrated Systems have demonstrated that microemulsions can be used to solubilize oxidized copper (Cu^{2+}) species, and based on these results they developed processes to remove post-breakthrough etch residues.[13,103] In a testing experiment, the microemulsion formulation containing 2% surfactant and water (17 mole/mole of surfactant) was able to rapidly solubilize Cu(II) complex at 2 mg/mL of $scCO_2$, equivalent to a solubility level approximately six orders of magnitude higher than that with the use of neat $scCO_2$ only. The solubility comparison was demonstrated visually using a view cell equipped with sapphire windows and on the basis of results from XPS analysis. For post-breakthrough residues removal, three blanket copper wafers supplied by International SEMATECH were first exposed to etch stop breakthrough gas (C_4F_8/O_2) to generate oxidized copper residues. Then the wafers were cleaned using either a commercial wet cleaning agent or two different CO_2-based microemulsions. The two wafers cleaned with the CO_2-based microemulsions were significantly higher in the copper level (indicating the removal of more residue) than the wafer cleaned with the commercial agent. The effectiveness of $scCO_2$-based microemulsions was also illustrated in post-ash cleaning processes, performed on patterned wafers containing two different low-κ spin-on materials (JSR 5109 from JSR, Tokyo and dense and porous SiLK from Dow Chemical, Michigan).[103] The results from the cross-sectional SEM imaging of the post-breakthrough JSR 5109 wafer as received, after water-in-CO_2 microemulsion cleaning, and after a process-of-record aqueous cleaning are compared in Figure 6.9. The as-received wafer sample shows granular residues on the floor and lower wall of the via and a white residue band halfway up the via wall. The aqueous cleaning removed all residues from the via floor but left some on the wall. The cleaning with the water-in-CO_2 microemulsion removed all residues without copper overetch and critical dimension (CD) loss in the low-κ dielectric.[103]

$ScCO_2$ mixed with small amounts of co-solvents were also investigated in the removal of post-etching residues from ultra low-κ films and patterned silicon surfaces.[102,104–108] An example was the cleaning of ultra low-κ film JSR 5109 wafer (nominal k value of 2.2), and the results are shown in Figure 6.10.[102] According to the SEM image, there was clearly the presence of globular residues on the sidewalls after ashing and etching. All of these post-ash residues in the wafer sample were removed upon a two-step sequence: the cleaning with $scCO_2$ and 3 vol.% co-solvent at 60°C and 2900 psi for 2 min, followed by a rinse in neat CO_2 at the same temperature and pressure for another 2 min (Figure 6.10). The CDs of both the trench and the via were preserved after the processing, and there was no apparent damage to the porous ultra low-κ film and the copper metal. Similarly, a formulation of tetramethylammonium hydroxide (TMAH), methanol, and water was used as co-solvent with $scCO_2$ to clean post-etch residues on the substrate of low-κ coral films.[104] The effect of process parameters, such as temperature, time, and co-solvent composition on the residue removal was investigated. The method demonstrated, on the basis of XPS and SEM analyses, a complete removal of photoresist film and etch residues by CO_2 with 12 wt% co-solvent (a 4:1 volumetric mixture of 25% TMAH in methanol and deionized water) at 70°C and

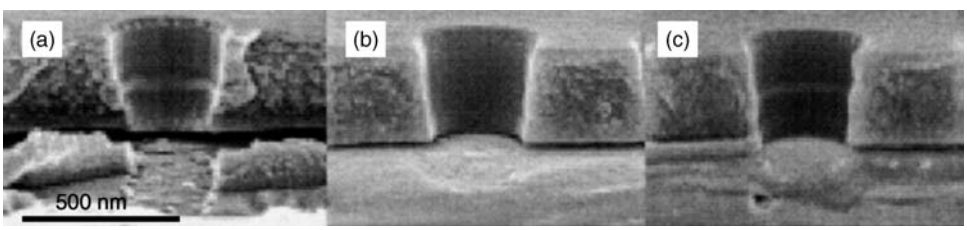


FIGURE 6.9 SEM images on the cross section in the post-breakthrough cleaning of patterned JSR 5109 structures: (a) as-received sample contains via bottom and sidewall residue; (b) cleaning with water-in-CO_2 microemulsion that completely removes residue; and (c) aqueous cleaning that removes residue from sidewalls but not the via bottoms. (From Wagner, M., DeYoung, J., Gross, S., Hatcher, Z., and Ma, C., in *ECS*, Pennington, NJ, 2003.)

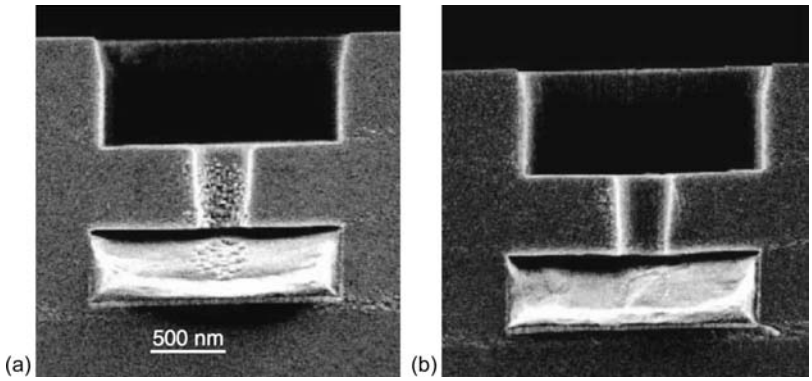


FIGURE 6.10 SEM images of dual damascene features before and after processing with $scCO_2$: (a) after ashing showing globular residue on sidewalls of via etched in ultra low-κ film and (b) after a two-step (cleaning and rinsing) process with $scCO_2$. (From Muscat, A.J., *Business Briefing: Global Semicond. Manuf. Technol.*, 2003.)

3000 psi. In addition, water or ozone has also been used in conjunction with pressurized CO_2 to provide a more aggressive environment for cleaning surfaces.[105,106]

Another important approach developed in the past decade is the in situ chelation method for dissolving metal species in $scCO_2$.[109] It was first reported by Laintz et al.,[110] who used fluorinated dithiocarbamate chelating agent to extract transition metal ions into $scCO_2$. After this initial report, the in situ chelation/metal extraction technique has been expanded to include other metal ions, especially lanthanides, actinides, uranium, and plutonium for the great potential in nuclear waste management. For the selective removal of metals, Glennon and co-workers[111] reported the selective extraction of gold into $scCO_2$ using fluorinated molecular baskets and thiourea ligands. In another demonstration, Kersch et al.[112] reported pilot-scale ligand-assisted extraction of toxic heavy metals from sewage fly ashes using Cyanex 302 (a trimethylphenyl-substituted monothiophosphinic acid) as an extractant. Soluble chelating agents, such as β-diketonates, crown ethers, dithiocarbamates, amines, hydroxamic acid, and organophosphates in $scCO_2$ have shown the effectiveness in interacting and removing metal ions. However, there have only been a few applications that exploit this approach for the cleaning of semiconductor devices.[113,114]For example, hexafluoroacetylacetone (hfacH) dissolved in $scCO_2$ was used to purposely remove contaminated silicon wafers of JSR 5109 porous SiOCH ($k=2.2$) with metallic copper (Cu(0)) and oxidized copper (CuO and Cu_2O).[113] The cleaning was performed in CO_2 + ethanol + hfacH at 61°C and 180 bar, and the evaluation on the residual contamination levels was based on the vapor-phase decomposition-atomic absorption spectroscopy. A cleaning efficiency of 98% was reached after only 1 min, and 99% shortly thereafter. In another example, Wang et al.[115] examined the effectiveness of the approach on chip resistors (electronic components commonly used in PC boards for surface mount technology). Most of the chip resistors today have wrap-around electroplated Ni or (Sn/Pb) terminations.[116] The removal of Ni and (Sn/Pb) from resistor surfaces after conventional electroplating is required to ensure the quality of solderability during termination soldering applications.[117,118] Four different cleaning procedures for removing the residued Ni or (Sn/Pb) ions were compared: (1) batch-cleaning using $scCO_2$; (2) batch-cleaning using $scCO_2$ infused with 0.1 M ethylenediaminetetraacetic acid disodium salt (EDTA); (3) flow-cleaning using deionized water; and (4) flow-cleaning using $scCO_2$. The results indicate that the flow-cleaning using $scCO_2$ was superior in the cleaning ability and the fastest in comparison with the other three methods. The addition of EDTA in $scCO_2$ in the batch-cleaning method did improve the cleaning ability. There was also evidence suggesting that higher pressure and temperature could significantly enhance the cleaning ability of $scCO_2$.[115]

6.3.3 Supercritical CO_2 in Etching

Etching is a surface treatment indispensable in semiconductor device fabrication. The selection of etchant is important in etching, depending on the semiconductor materials involved and on other materials present, such as metals, oxides, and photoresist. The purpose of etching can be either to clean the surfaces and to remove mechanical damage caused by sawing operations, to reveal defects and impurity precipitates in crystals, or to obtain special shapes (mostly on a small scale) as required for the proper functioning of devices.[119] There are two stages in the etching of semiconductors: oxidation and the subsequent dissolution of the oxidation products. Traditional etching processes are wet, where the material is dissolved in the isotropic or anisotropic way when immersed in a chemical solution. In dry etching, the material is sputtered or dissolved using reactive ions or a vapor-phase etchant. Isotropic etchants (for example, hydrofluoric acid (HF)), available for oxide, nitride, aluminum, polysilicon, gold, and silicon, attack the material being etched at the same rate in all directions. Anisotropic etchants (for example, potassium hydroxide) attack the silicon wafer at different rates in different directions, and thus allowing more control of the shapes produced. Most common surface layers etched today in the new generation of microelectronic devices are SiO_2, Si_3N_4, and copper. The traditional etching techniques are often ineffective, damaging toward new interconnect materials, environmentally unacceptable, etc. $ScCO_2$ is being considered as an alternative in the etching of silicon oxide-type materials used as dielectrics or sacrificial layers in microelectronic fabrication, and for the removal of residues formed during plasma etching steps. It would eliminate the need for subsequent rinsing and drying steps and prevent chemical modification or structural damage to materials in semiconductor devices. In this regard, $scCO_2$-based nonaqueous HF etchant solutions have been developed for dry-etch processing of microelectronic devices.[120,121] Generally, the cleaning fluid comprises CO_2, HF, and a Lewis base (in particular, an amine such as pyridine, poly(vinylpyridine), or triethyl amine).[120,121] The high-pressure etching apparatus compatible with HF has been designed and implemented (Figure 6.11),[121] in which all wetted parts are made of corrosion-resistant alloys (Hastelloy, Monel, or Ni 200). The CO_2-based formulations containing HF/pyridine complex were tested for etching on dual damascene wafers. The first set of experiments were designed to investigate the effect of HF concentration, temperature, and CO_2 pressure on the dissolution of SiO_2 thin film, and the amount of SiO_2 etched away was found to increase with increases in these parameters. The average rate ranged from 5 to 40 Å/min, depending on the experimental conditions.[121]

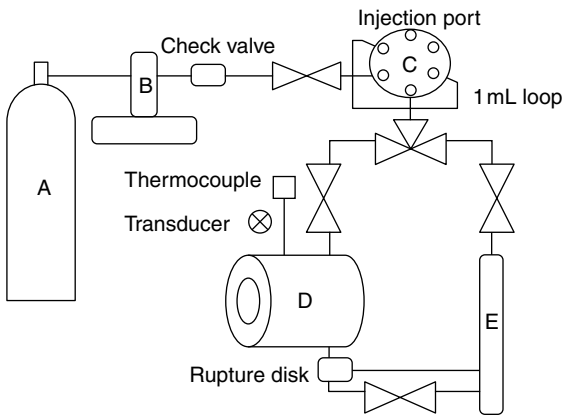


FIGURE 6.11 A setup for high-pressure CO_2/hydrofluoric acid etching: (a) CO_2 cylinder; (b) syringe pump; (c) injection valve with 1 mL equilibration loop; (d) 10 mL etching chamber; and (e) PVC evacuation port with 1 M NaOH. (From Jones III, C.A., Yang, D., Irene, E.A., Gross, S.M., Wagner, M., DeYoung, J., and DeSimone, J.M., *Chem. Mater.*, 15, 2867, 2003.)

The application of this etchant formulation was also convenient for post-etch residue removal in current back-end-of-line or BEOL (operations performed after first metallization on the semiconductor wafer) cleaning in the semiconductor industry.[121] The test was performed on dual damascene structures containing Silox as the dielectric layer and Si_3N_4 as the etch stop layer. The SEM imaging of the cross-section of an "as-received" sample and a treated sample with the CO_2-based etchant solution yielded the results shown in Figure 6.12. The etch stop layer obscured by etch residues in the as-received sample reappears after exposure to an excess etchant solution. A problem was some loss in the CD, though it could be avoided by a careful control of the etching rate.[121] In another recent study,[122] the removal of native SiO_2 and GeO_2 layers from SiGe films (15%–30% Ge) using $scCO_2$ with aqueous HF was explored.

Researchers have also demonstrated that $scCO_2$-based systems are also well suited for MEMS, including the etching of sacrificial oxides for the release of mechanical parts, the cleaning of their surfaces, and the maintaining clean environment before passivation.[123–126] MEMS are small integrated devices that combine electrical and mechanical components, often manufactured by surface micro-machining techniques.[123,127] The manufacturing process involves depositing multiple thin layers of material on silicon wafer, patterning, and etching. The final step is the removal of sacrificial oxide layer which is done usually through wet etching with an aqueous HF solution, followed by rinsing in deionized water and drying. The use of these standard release methods has resulted in challenges such as inefficacy in complete residue removal from the surface, and sticking and collapsing of microstructures due to capillary forces. $ScCO_2$ drying has already proven effective for high yields with microstructures fabricated beneath thin-film shells and for preventing sticking and collapsing.[123–126]

The growing interest in using copper as the interconnect metal and the drive toward miniaturization in the microelectronics industry present tough challenges to the patterning of copper using conventional dry-etch techniques and no tolerance for defects as the feature size reduces. In this regard, $scCO_2$-based systems and chemical mechanical planarization (CMP) process have emerged as revolutionizing technologies for implementing copper in submicron semiconductor devices.[128–131] CMP is a process of smoothing and planing surfaces with a combination of chemical and mechanical forces, a hybrid of chemical etching and free abrasive polishing. It is used to make metal and dielectric layers on silicon substrates smooth and defect-free with vertical dimension control. In copper CMP, the typical chemical slurry used to remove copper from wafer contains chemical etchants (oxidants), chelating agents, buffers, abrasive particles, passivating agents, and co-solvent(s). Most of the current processes use water and chlorinated organic solvents for the CMP slurry, causing technical (such as the incompatibility of aqueous slurries with porous low-κ interlayer dielectric materials) and environmental difficulties. The inclusion of CO_2 in CMP slurries helps to reduce waste streams, easily separated from other ingredients

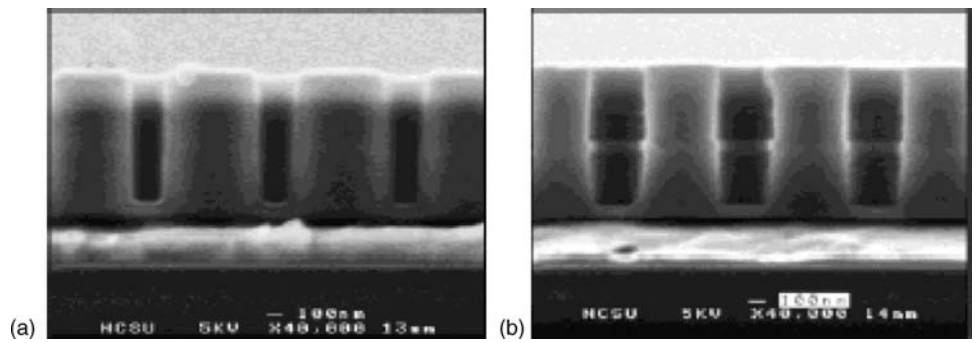


FIGURE 6.12 SEM images of the two cross sections showing dual damascene structures containing SiO_x as the dielectric layer and Si_3N_4 as the etch stop layer: (a) the "as-received" sample post-dielectric etch and (b) treated with excess CO_2-based etchant solution (critical dimension loss observed). (From Jones III, C.A., Yang, D., Irene, E.A., Gross, S.M., Wagner, M., DeYoung, J., and DeSimone, J.M., *Chem. Mater.*, 15, 2867, 2003.)

of the slurries, and to improve the compatibility with porous low-κ interlayer dielectric materials. Douglas and Templeton[132] have issued two patents on the removal of metal atoms from wafer surfaces using a chelating agent dissolved in $scCO_2$. Bessel et al.[131] demonstrated the etching of Cu coupons in a mixture of the oxidant ethyl peroxydicarbonate and different β-diketonate chelating molecules dissolved in $scCO_2$ at 40°C and 214 bar. Their effort resulted in the first report on the development of a $scCO_2$-based CMP process. They found that the removal rate with the use of hexafluoroacetylacetone (hfacH) was 158 copper layers per minute or approximately three times greater than that with chelating molecules containing less fluorine, and they attributed the improvement to the solubility of the Cu-bearing complexes in CO_2. Recently, Xie et al.[133] reported a proof-of-concept process for the etching of copper and copper oxide from a silicon surface using hfacH dissolved in $scCO_2$ at 40°C–60°C and 100–250 bar. Copper islands with nominal concentrations of the order of 10^{16} atoms/cm^2 were chosen for the study since the concentration range is most relevant to wafer surface cleaning applications in the semiconductor industry. A two-step process of oxidation and etching with hfacH dissolved in $scCO_2$ removed Cu on Si. XPS results indicated the presence of $Cu(I)_2O$ in those islands due to the air exposure before etching. Cu(I) was partially oxidized using aqueous 30 vol.% H_2O_2 or UV–Cl_2 gas phase, forming shells of Cu(II)O or $Cu(II)Cl_2$, respectively. Both Cu(II)O and $Cu(I)_2O$, as well as copper chlorides, were chemically etched using hfacH dissolved in $scCO_2$. Nucleophilic attack of Cu(II) centers by hfacH formed copper(bis-hexafluoroacetylacetonate), $Cu(hfac)_2$ and water, or the monohydrate $Cu(hfac)_2 \cdot H_2O$, which was soluble in $scCO_2$. It was proposed that the $Cu(hfac)_2 \cdot H_2O$ byproduct oxidized $Cu(I)_2O$ to Cu(II), allowing attack and etching by hfacH.[133] These findings were unprecedented in the oxidation and removal of an elemental metal film in CO_2.

6.3.4 Supercritical CO_2 in Drying

There is a long line of process steps in the manufacturing of microelectronic structures, and drying is considered a step of particular importance. The drying, which is done at the end of almost every wet process sequence, contributes tremendously to the process performance or may even define the outcome. The processing of an increasing number of 300-mm wafers, the vertical integration, and the shrinking of CDs to less 100-nm nodes, all create new challenges to the drying process. Accordingly, height:width aspect ratios of 50–60:1 have become standard for deep trenches, and advanced applications are already at ratios of up to 100:1. These complex wafer geometries and structures have led to very mechanically fragile structures, redefining some of the major wafer drying performance requirements. Pattern collapse (deformation or bending) of structures is a phenomenon related to the surface tension of rinse solution, and it tends to enhance with increasing surface tension and aspect ratio and decreasing spacing. Two different approaches have been employed during the last decade to overcome this problem. The first was to increase the material stiffness by hardening the structures and improving their mechanical properties and therefore achieving better performance.[134–136] The second was to focus on minimizing or eliminating the surface tension during the drying stage. Low surface tension solvents have been tested as rinse liquids, but the choice has been limited by the chemical compatibility with the photoresist resin.[137] Freeze-drying of a rinse liquid (for example, *t*-butanol) has sometimes been effective in preventing pattern collapse.[138] The solvent to be freeze-dried must have a low volume change upon freezing and a melting point close to room temperature, and must not strip the resist. These drying techniques are approaching their limits, generally lack of the inherent potential to meet the demands of future technologies.

It is acknowledged that the best solution to this issue is the application of supercritical fluid drying, primarily with the use of $scCO_2$. The absence of surface tension in supercritical fluids has been widely exploited in the removal of solvents, particularly from aerogels,[139] without structure collapse. Successful results were obtained when an alcohol was employed for rinsing, followed by $scCO_2$.[140,141] Alcohol readily dissolves in CO_2 and is also easy to handle. This method was used by Namatsu et al.[140,141] to dry fine silicon patterns without any collapse. For resist-pattern drying, pattern deformation was found to be severe after supercritical drying because of water contamination.[141] It was shown that the presence of

moisture in the dryer chamber could cause the acrylate-type resist to swell, because H_2O molecules became seeds for the release of gases from inside the resist film. In order to avoid the resist-film swelling, it was necessary to eliminate water contamination by injecting $scCO_2$ into the chamber. The method was applied to 20-nm wide patterns of aspect ratios as high as 7.5 without any collapse.[141] For resist systems designed to work with aqueous-based developers, the rinse water cannot simply be replaced with CO_2, because water hardly dissolves in CO_2. Two alternative methods of supercritical drying by integrating a surfactant were developed to overcome this problem. In the first method, liquid CO_2 containing CO_2-philic surfactant (such as fluoroether carboxylate) replaces the water directly, followed by liquid CO_2 exchange. In the second method, the rinse water is first replaced with a surfactant in an organic liquid, such as hexane, followed by liquid CO_2 exchange. In both ways, the liquid CO_2 is subsequently converted to its supercritical state to effect the drying operation. The surfactant in the second method facilitates the incorporation of the aqueous phase into micellar microdomains in the organic phase. The best surfactant used by Namatsu[142] was a mixture of sorbitan monolaurate with a small amount of polyethylene glycol mono-4-nonylphenyl ether. Golfarb et al.[143] used a different surfactant, sodium dioctylsulfocuccinate (AOT), and demonstrated that the structures could be perfectly preserved without collapsing for aspect ratios up to 6.8.

The CO_2-based drying was also found to be effective in removing water from within the pores of ultra low-κ films in dual damascene features.[102] Water is usually introduced into organosilicate low-κ layers during ashing of photoresist and the subsequent cleaning, which preferentially attacks the silicon–carbon bonds. The presence of water induces the formation of silanol (Si-OH) groups, which raise the k value typically above 3. In order to repair the damage and reduce the dielectric constant, the Si-OH groups must be removed using a condensation reaction. In this regard, Muscat and co-workers[102,144,145] treated pMSQ films with $scCO_2$ containing alcohol and carboxylic acid as co-solvents. Alcohols as well as acetic acid are effective in removing H-bonded Si-OH groups, and isopropanol, 1-propanol, and 1-butanol offer the best performance at the lowest vapor pressure. These alcohols are polar molecules, which increase the solubility of water in $scCO_2$, doubling the removal efficiency. The FTIR results indicated that about 25% of the hydrogen-bonded Si-OH groups were removed after processing pMSQ films in neat $scCO_2$ at 55°C and 189 atm. The introduction of 5 vol.% 1-propanol as co-solvent into $scCO_2$ at 49°C and 279 atm allowed the removal of 60% of the water in just 2 min. The vacuum annealing to 200°C accomplished the same objective of removing about half of the hydrogen-bonded Si-OH groups. Another approach to remove water and restore the k value of the film was to simply chemically react the damaged sites with a silicon-based reagent, such as trimethylchlorosilane (TMCS).[146,147] The treatment of an etched and ashed ultra low-κ MSQ film with 1 vol.% TMCS in $scCO_2$ at 50°C and 299 atm for 2 min removed the water from not only hydrogen-bonded but also isolated Si-OH groups.

The CO_2-based drying is also particularly useful in the removal of water and other contaminants from MEMS-based devices.[125,148–150] The procedure was found to be convenient in preventing stiction and collapsing which often occur in MEMS devices following their final release etch. Researchers at the Sandia National Laboratory demonstrated the effectiveness of CO_2-based drying and cleaning on singly clamped cantilever beam components (parts in microengine device).[125] In that study, a combination of $scCO_2$–methanol drying was used to decrease the interfacial adhesional energy by three orders of magnitude, and subsequently to enhance the yield of microengines by a factor of 3.

6.4 Processing Equipment

Laboratory scale equipment has mostly been used in the implementation of $scCO_2$-based processes.[11,51,52,66,151–153] In each process, specific CO_2-based formulations were first optimized usually on wafer fragments in 5–25 mL high-pressure treatment vessels. Liquid CO_2 supplied from either a cooled bulk tank or a cylinder is fed into a high-pressure pump to attain the desired supercritical regime. Chemicals are introduced into the pressurized liquid CO_2, and the mixture is typically stored in vessels capable of delivering the desired quantities and pressures into the treatment vessel at high flow.

For the final rinsing step, the treatment vessel is purged with neat CO_2 at the same processing pressure to avoid deposit of additives on the wafers already treated. CO_2 is then separated from chemicals into the vessel known as separator for recovery and recycle, and byproducts of the process are removed. The processing efficacy is usually determined by analyzing the treated sample with established surface characterization techniques, such as FTIR, SEM, Auger and NRA, XPS, etc.

The LANL-SCORR process for the removal of photoresist as well as post-etching and other treatment residues during the patterning of semiconductor wafers is considered as one of the original research systems designed around the use of $scCO_2$.[11,51,52] It is a process that holds the promise of enabling the global integrated circuit industry to achieve its goal of producing increasingly higher-density microchips than those that can be made today. Because of its advanced cleaning process, SCORR is also considered to be compatible with the latest low-κ and smaller ($<0.18\ \mu m$) dimensions necessary to advance the industry into the future. Patented by the Laboratory, the technology is now available to integrated circuit makers through a license to SC Fluids, Inc. (Nashua, N.H.), and the company is partnering with IBM to commercialize the technology. The next challenge is to scale up to a full 200-mm wafer process. In fact, supercritical fluid technologies based on CO_2 are more convenient to scale up to industrial processes because of their attractive properties and their relatively low costs. CO_2 is usually mined or recovered from alcohol fermentation plants or lime furnaces used in cement manufacturing, and it can also be recovered as a waste gas from coal, oil, and gas power plants.[93] It offers many substantial benefits, both environmentally and economically. According to SC Fluids, Inc., the estimate is that the CO_2 process uses 99% less chemicals than the standard wet bench cleaning.[93,153–157] The equipment that the company is designing will replace both a plasma asher and the solvent wet bench where wafer cleaning is performed, and the cost is expected to be less than the combination of the pieces that it replaces.[93,153–157] The company argues that the major economic benefits will come from the lower cost of running the equipment because it skips the rinsing and drying steps and eliminates the costs associated with the disposal of waste solvents. Technically, it has already been shown at the laboratory scale that the CO_2-based formulations can be similarly recreated in 200-mm single-wafer processors as long as the processing pressure, temperatures, and time are the same.[13,93,153–156]

For the eventual implementation of the SCORR process, SC Fluids, Inc. and partners enabled the tool development of a fully automatic system called "ARROYO" (Figure 6.13).[156,157] It has been successfully operated for process applications, with the capacity to process 50 wafers in two 25-wafer lots. Similar to the SCORR process, the system consists of a pressure vessel that can treat either 150- or 200-mm wafers and a fluid delivery/separator system. CO_2 can be delivered to the pressure vessel in gas, liquid, or supercritical state at any appropriate temperature and pressure. Precise concentrations of co-solvents or surfactant mixtures with $scCO_2$ can also be delivered to the pressure vessel. Surfactants can be used in a separate process to remove particles other than photoresist from wafers. After the wafers are cleaned, CO_2 is reverted to a gas form and separated from the co-solvents or surfactants, as well as the resist debris from the wafers. The recovered CO_2 can be recycled, so can the co-solvents, and the debris is disposed of as waste. Recently, the company sold the first commercial 300-mm $scCO_2$ wafer processing system to a major East Coast integrated circuit manufacturer.[156,157] The fully automated 300-mm ARROYO system is for single wafer, cassette-to-cassette. The system is designed to clean photoresist and for post-etch residue removal from low-κ films with <130 nm features, high energy/high dose implanted resist removal, photoresist image collapse prevention, and NGL Photo mask cleaning and particle removal. It has successfully passed a SEMI S2-0302 audit for environmental, health and safety.

According to a program director at SEMATECH, a consortium of semiconductor manufacturing companies based in Austin-Texas, the $scCO_2$ technology developed at LANL, if proven, could be a tremendous plus for the industry, as well as for the environment.[93] "We are still in the very early stages on this technology", the director cautions. "One of the characteristics that could make SCORR a win–win situation is not just its environmental friendliness, but also its potential."[93] "If we can develop this technology and build it in as were building a new facility, that would be the best opportunity to make an impact from an economic perspective. In any event, going to more benign resist-removal processes,

FIGURE 6.13 The 300-mm Arroyo system for cleaning semiconductor chips. (From Rothman, L., *Annual Retreat*, 2003. Available online http://www.gscn.net/event/images/Speech-6.PDF.)

processes that are safer for the environment and for our workers and that use fewer natural resources, would be a tremendous benefit for the industry and the world as a whole."[93]

6.5 Conclusions and Perspectives

It is apparent that over the past few years, exciting technological progress has been made in the design of $scCO_2$-based formulations and the development of equipment for the implementation of the supercritical fluid processing to address many of the semiconductor cleaning challenges. This new technology outperforms conventional processes in many areas, including waste minimization, water use, energy consumption, worker safety, reduction in the number of process steps, compatibility with submicron features, and costs. Several successful cleaning applications have already been accomplished for the removal of photoresist, residues, and particles from the smallest features in integrated circuits. Significant successes have also been demonstrated in the etching of metal and oxides films, and the drying of pores in ultra low-κ films in dual damascene features and MEMS-based devices without pattern collapsing in the structures. Because of the weak solvation power of $scCO_2$ alone, modifiers such as co-solvents, surfactants, chelating agents, and/or chemical reactants are used with CO_2 to enhance the solubility performance in semiconductor processing.

The experiments have generally been performed in a closed loop treatment system, in which liquid CO_2 is introduced and brought to the supercritical state and combined with modifier(s), and the wafer

surface is exposed to the resulting supercritical fluid. Upon depressurization, the CO_2 is easily recovered and recycled after the chemicals and process byproducts are removed, leaving behind the wafers dry and residue free. The process can be made selective toward the materials it cleans using appropriate pressure, temperature, and modifier(s). The properties of relatively high and tunable density, low viscosity, and low surface tension characteristic of $scCO_2$ are the keys to the cleaning effectiveness. These properties make it possible to overcome some of the major technological challenges facing the semiconductor industry in the drive toward miniaturization, such as the requirement for high mass transport, penetration of small features, wetting of advanced low-κ materials, and patterning of devices with high aspect ratio vias and trenches.

Despite the significant progress, the field is still considered to be in early stages. For example, some of the current processes still need to be optimized and improved by investigating and understanding effects of various experimental conditions. More selections of soluble, nontoxic, and efficient modifiers for $scCO_2$ are also needed. Other areas of need are the knowledge on cleaning mechanisms and the development of practical equipment and technical approaches for the new cleaning systems. The incorporation of the $scCO_2$-based cleaning process into the design and manufacture of next-generation semiconductor devices is obviously challenging. To achieve higher performance and more functionality, circuits patterns will become even finer, from 90 to 65 nm and even 45 nm, and the number of interconnect layers will increase as well. These along with the use of new materials and the ongoing development of design technologies for three-dimensional spatial structures, such as MEMS, demand substantial improvements in cleaning processes. Furthermore, the use of the technology at the industrial scale is also a very important issue, for which a likely solution is the collaboration between industry, government, and academia. In any case, the future is bright for the application of supercritical fluid technology in semiconductor industry.

References

1. Van Zant, P. *Microchip Fabrication*. 4th ed., New York: McGraw Hill, 2000.
2. Kern, W. "Overview and Evolution of Semiconductor Wafer Contamination and Cleaning Technology." In *Handbook of Semiconductor Wafer Cleaning Technology—Science, Technology, and Applications*, edited by W. Kern, Noyes: Andrew Publishing, 1993.
3. Hattori, T. "Contamination Control: Problems and Prospects." *Solid State Technol.* 33 (1990): 1.
4. Williams, E. D., R. U. Ayres, and M. Heller. "The 1.7 Kilogram Microchip: Energy and Material Use in the Production of Semiconductor Devices." *Environ. Sci. Technol.* 36 (2002): 5504.
5. *International Technology Roadmap for Semiconductors*. San Jose: Semiconductor Industry Association, 2003, Available at: http://public.itrs.net
6. Maier, G. "Low Dielectric Constant Polymers for Microelectronics." *Prog. Polym. Sci.* 26 (2001): 3.
7. McHardy, J., and S. P. Sawan. *Supercritical Fluid Cleaning—Fundamentals, Technology, and Applications*. Westwood: Noyes Publications, 1998.
8. Kanegsberg, B., and E. Kanegsberg. *Handbook for Critical Cleaning*. Boca Raton: CRC Press, 2001.
9. Mount, D. J., L. B. Rothman, R. J. Robey, and M. K. Ali. "The Technology behind Cleaning with Supercritical Fluids." *Solid State Technol.* 7 (2002): 103.
10. Weibel, G. L., and C. K. Ober. "An Overview of Supercritical CO_2 Applications in Microelectronic Processing." *Microelectron. Eng.* 65 (2003): 145.
11. King, J. W., and L. L. Williams. "Utilization of Critical Fluids in Processing Semiconductors and Their Related Materials." *Curr. Opin. Solid State Mater. Sci.* 7 (2003): 413.
12. Jones, C. A. III., A. Zweber, J. P. DeYoung, J. B. McClain, R. Carbonell, and J. M. DeSimone. "Applications of 'Dry' Processing in the Microelectronics Industry Using Carbondioxide." *Crit. Rev. Solid State Mater. Sci.* 29 (2004): 97.
13. Case, C., B. O. C. Edwards, and J. McClain. "Micell Integrated Systems, Developing Supercritical Carbon Dioxide Processing in Microelectronics Applications." *MICRO,* January/February, (2004).
14. Moslehi, B. "Examining the Future of Wafer Cleaning." *MICRO,* May, (2004).

15. Beckman, E. J. "Green Chemical Processing Using CO_2." *Ind. Eng. Chem. Res.* 42 (2003): 1598.
16. Wells, S. L., and J. DeSimone. "CO_2 Technology Platform, an Important Tool for Environmental Problem Solving." *Angew. Chem. Int. Ed.* 40 (2001): 518.
17. O'Murchu, C., A. Mathewson, and E. Francais. "Effects of Supercritical CO_2 on the Electrical Characteristics of Semiconductor Devices." *Electrochem. Soc. Proc.* 26 (2001): 305.
18. Falconer, J. L., S. D. Bischke, and G. H. Hanna. "Electron-Enhanced CO_2 Adsorption and Stabilization on Aluminum Films." *Surf. Sci.* 131 (1983): 455.
19. Russick, E. M., G. A. Poulter, C. L. J. Adkins, and N. R. Sorensen. "Corrosive Effects of Supercritical Carbon Dioxide and Cosolvents on Metals." *J. Supercrit. Fluids* 9 (1996): 43.
20. Behner, H., W. Spiess, G. Wedler, and D. Borgmann. "Interaction of Carbon Dioxide with Fe(110), Stepped Fe(110) and Fe(111)." *Surf. Sci.* 175 (1986): 276.
21. Collins, A. C., and B. M. W. Trapnell. "CO_2 Chemisorption on Evaporated Metal Films." *Trans. Faraday Soc.* 53 (1957): 1476.
22. Habraken, F. H. P. M., E. Ph. Kieffer, and G. A. Bootsma. "A Study of the Kinetics of the Interaction of O_2 and N_2O with a Cu(111) Surface and the Reaction of CO with Adsorbed Oxygen Using AES, LEED Ellipsometry." *Surf. Sci.* 83 (1979): 45.
23. Krause, J., D. Borgmann, and G. Wedler. "Photoelectron Spectroscopic Study of Adsorption of Carbon Dioxide Cu(110) and Cu(111)/K—Compared with the Systems Fe(110)CO_2 and Fe(110)/K+CO_2." *Surf. Sci.* 347 (1996): 1.
24. Sherman, R., D. Hirt, and R. Vane. "Surface Cleaning with the Carbon Dioxide Snow Jet." *J. Vac. Sci. Technol.* 12 (1867): 1994.
25. Darr, J. A., and M. Poliakoff. "New Directions in Inorganic and Metalorganic Coordination Chemistry in Supercritical Fluids." *Chem. Rev.* 99 (1999): 495.
26. Quinn, E. L., and C. L. Jones. *Carbon Dioxide*. New York: Reinhold, 1936.
27. Bunker, C. E., H. W. Rollins, and Y.-P. Sun. "Fundamental Properties of Supercritical Fluids." In *Supercritical Fluid Technology in Materials Science and Engineering: Synthesis, Properties, and Applications*, edited by Y.-P. Sun, 1. New York: Marcel Dekker, 2002.
28. Sun, Y.-P., H. W. Rollins, J. Bandara, M. J. Meziani, and C. E. Bunker. "Preparation and Processing of Nanoscale Materials by Supercritical Fluid Technology." In *Supercritical Fluid Technology in Materials Science and Engineering: Synthesis, Properties, and Applications*, edited by Y.-P. Sun, 491. New York: Marcel Dekker, 2002.
29. Kamlet, M. J., J. L.-M. Abbound, and R. W. Taft. "The Solvatochromic Comparison Method. 6. The .pi.* Scale of Solvent Polarities." *J. Am. Chem. Soc.* 98 (1977): 6027; Kamlet, M. J., T. N. Hall, J. Boykin, and R. W. Taft. "Linear Solvation Energy Relationships. 6. Additions to and Correlations with the .pi.* Scale of Solvent Polarities." *J. Org. Chem.* 44 (1979): 2599.
30. Dong, D. C., and M. A. Winnik. "The Py Scale of Solvent Polarities-Solvent Effects on the Vibronic Fine Structure of Pyrene Fluorescence and Empirical Correlations with Et-Value and Y-Value." *Photochem. Photobiol.* 35 (1982): 17; Dong, D. C., and M. A. Winnik. "The Py Scale of Solvent Polarities." *Can. J. Chem.* 62 (1984): 2560.
31. Rettig, W. "Charge Separation in Excited-states of Decoupled Systems—TICT Compounds and Implications Regarding the Development of New Laser-Dyes and the Primary Processes of Vision and Photosynthesis." *Angew. Chem. Int. Ed. Eng.* 25 (1986): 971.
32. Sun, Y.-P., C. E. Bunker, and N. B. Hamilton. "Py scale in Vapor-phase and in Supercritical Carbon-dioxide Evidence in Support of a 3-Density-Region Model for Solvation in Supercritical Fluids." *Chem. Phys. Lett.* 210 (1993): 111.
33. Sun, Y.-P., and C. E. Bunker. "Quantitative Spectroscopic Investigation of Enhanced Excited State Complex Formation in Supercritical Carbon Dioxide under Near-Critical Conditions: Inconsistency between Experimental Evidence and Classical Photophysical Mechanism." *J. Phys. Chem.* 99 (1995): 13786.
34. Sun, Y.-P., and C. E. Bunker. "Solute and Solvent Dependence of Intermolecular Interactions in Different Density Regions in Supercritical Fluids—a Generalization of the 3-Density-Region Solvation Mechanism." *Ber. Bunsen-Ges. Phys. Chem.* 99 (1995): 976.

35. Weber, M., and M. C. Thies. "Understanding the RESS Process." In *Supercritical Fluid Technology in Materials Science and Engineering: Synthesis, Properties, and Applications*, edited by Y.-P. Sun, 387. New York: Marcel Dekker, 2002.
36. Andrew, D., B. T. Des Islet, A. Margaritis, and A. C. Weedon. "Photo-Fries Rearrangement of Naphthyl Acetate in Supercritical Carbon Dioxide—Chemical Evidence Solvent–Solute Clustering." *J. Am. Chem. Soc.* 117 (1995): 6132.
37. Hrnjez, B. J., A. J. Mehta, M. A. Fox, and K. P. Johnston. "Photodimerization of Isophorone in Supercritical Trifluoromethane and Carbon Dioxide." *J. Am. Chem. Soc.* 111 (1989): 2662.
38. Kimura, Y., and Y. Yoshimura. "Chemical Equilibrium from the Gaseous to Liquid States: Solvent Density Dependence of the Dimerization Equilibrium of 2-Methyl-2-Nitrosopropane in Carbondioxide, Chlorotrifluoromethane, and Trifluoromethane." *J. Chem. Phys.* 96 (1992): 3085.
39. Weinstein, R. D., A. R. Renslo, R. L. Danheiser, J. G. Harris, and J. W. Tester. "Kinetic Correlation of Diels–Alder Reactions in Supercritical Carbon Dioxide." *J. Phys. Chem.* 100 (1996): 12337.
40. Isaacs, N. S., and N. J. Keating. "The Rates of a Diels–Alder Reaction and Supercritical Carbondioxide." *J. Chem. Soc. Chem. Commun.* (1992): 876.
41. Bunker, C. E., H. W. Rollins, J. R. Gord, and Y.-P. Sun. "Efficient Photodimerization Reaction of Anthracene in Supercritical Carbon Dioxide." *J. Org. Chem.* 62 (1997): 7324.
42. Vasapollo, G., L. Longo, L. Rescio, and L. Ciurlia. "Innovative Supercritical CO_2 Extraction of Lycopene from Tomato in the Presence of Vegetable Oil as Cosolvent." *J. Supercrit. Fluids* 29 (2004): 87.
43. Ohashi, K., and K. Tatenuma. "Extraction Behavior of Gallium(III) with 2-Methyl-5-Hexyloxymethyl-8-Quinolinol and 5-Hexyloxymethyl-8-Quinolinol from Weakly Acidic Solution into Supercritical CO_2 and Selective Separation of Gallium(III) from Aluminum(III)." *Chem. Lett.* 11 (1997): 1135.
44. Gamlieli-Bonshtein, I., E. Korin, and S. Cohen. "Selective Separation of cis-trans Geometrical Isomers of Beta-Carotene via CO_2 Supercritical Fluid Extraction." *Biotechnol. Bioeng.* 80 (2002): 169.
45. Matsuyama, H., A. Yamamoto, H. Yano, T. Maki, M. Teramoto, K. Mishima, and K. Matsuyama. "Effect of Organic Solvents on Membrane Formation by Phase Separation with Supercritical CO_2." *J. Membr. Sci.* 204 (2002): 81.
46. Eckert, C. A., D. Bush, J. S. Brown, and C. L. Liotta. "Tuning Solvents for Sustainable Technology." *Ind. Eng. Chem. Res.* 39 (2000): 4615.
47. Kendall, J. L., D. A. Canelas, J. L. Young, and J. M. DeSimone. "Polymerizations in Supercritical Carbon Dioxide." *Chem. Rev.* 99 (1999): 543.
48. Kompella, U. B., and K. Koushik. "Preparation of Drug Delivery Systems using Supercritical Fluid Technology." *Crit. Rev. Ther. Drug Carrier Syst.* 18 (2001): 173.
49. Pathak, P., M. J. Meziani, and Y.-P. Sun. "Supercritical Fluid Technology for Enhanced Dry Delivery." *Expert Opin. Drug Delivery* 2, no. 4 (2005): 747–61.
50. Xu, C., D. W. Minsek, J. F. Roeder, M. B. Korzenski, and T. H. Baum. Supercritical cleaning of semiconductor substrates. US Patent Appl. 20030125225, 2003.
51. Rubin, J. B., L. B. Davenhall, C. M. V. Taylor, L. D. Sivils, and T. Pierce. "Carbon Dioxide-Based Supercritical Fluids as IC Manufacturing Solvents." In *IEEE International Symposium on Electronics and the Environment*, Danvers, MA, 1999.
52. Biberger, M. A., W. H. Mullee, and P. E. Schilling. Removal of photoresist and residue from substrate using supercritical carbon dioxide process. Supercritical Systems Inc., WO0133613, 2001.
53. Harrison, K., J. Goveas, K. P. Johnston, and E. A. O'Rear. "Water-in-Carbon Dioxide Microemulsions with a Fluorocarbon-Hydrocarbon Hybrid Surfactant." *Langmuir* 10 (1994): 3536.
54. Johnston, K. P., K. Harrison, M. J. Clarke, S. M. Howdle, M. P. Heitz, F. V. Bright, C. Carlier, and T. W. Randolph. "Water in Carbon Dioxide Microemulsions: An Environment for Hydrophiles Including Proteins." *Science* 271 (1996): 624.
55. Eastoe, J., Z. Bayazit, S. Martel, D. C. Steyler, and R. K. Heenan. "Droplet Structure in a Water-in-CO_2 Microemulsion." *Langmuir* 12 (1996): 1423.

56. Keiper, J. S., J. A. Behles, T. L. Bucholz, R. Simhan, and J. M. DeSimone. "Self-Assembly of Phosphate Fluorosurfactants in Carbon Dioxide." *Langmuir* 20 (2004): 1065.
57. Ryoo, W., S. E. Webber, and K. P. Johnston. "Water-in-Carbon Dioxide Microemulsions with Methylated Branched Hydrocarbon Surfactants." *Ind. Eng. Chem. Res.* 42 (2003): 6348.
58. Johnston, K. P. "Block Copolymers as Stabilizers in Supercritical Fluids." *Curr. Opin. Colloid Interface Sci.* 5 (2000): 351.
59. DeSimone, J. M., and J. S. Keiper. "Surfactants and Self-Assembly in Carbon Dioxide." *Curr. Opin. Solid State Mater. Sci.* 5 (2001): 333.
60. Ye, X. G., and C. M. Wai. "Making Nanomaterials in Supercritical Fluids: A Review." *J. Chem. Educ.* 80 (2003): 198.
61. Taylor, D. K., J. S. Keiper, and J. M. DeSimone. "Polymer Self-Assembly in Carbon Dioxide." *Ind. Eng. Chem. Res.* 41 (2002): 4451.
62. Woods, H. M., M. M. C. G.Silva, C. Nouvel, K. M. Shakesheff, and S. M. Howdle. "Materials Processing in Supercritical Carbon Dioxide: Surfactants, Polymers and Biomaterials." *J. Mater. Chem.* 14 (2004): 1663.
63. Loeker, F., P. C. Marr, and S. M. Howdle. "FTIR Analysis of Water in Supercritical Carbon Dioxide Microemulsions Using Monofunctional Perfluopolyether Surfactants." *Colloid Surf. A* 214 (2003): 143.
64. Lee, C. T., P. A. Psathas, K. J. Ziegler, K. P. Johnston, H. J. Dai, H. D. Cochran, Y. B. Melnichenko, and G. D. Wignall. "Formation of Water-in-Carbon Dioxide Microemulsions with a Cationic Surfactant: A Small-Angle Neutron Scattering Study." *J. Phys. Chem. B* 104 (2000): 11094.
65. Bowling, A., B. Kirkpatrick, T. Hurd, L. Losey, and P. Matz. "Future Challenges for Cleaning in Advanced Microelectronics." *Solid State Phenom.* 92 (2003): 1.
66. Bok, E., D. Kelch, and K. S. Schumacher. "Supercritical Fluids for Single Wafer Cleaning." *Solid State Technol.* 35 (1992): 117.
67. DeForest, W. S. *Photoresist, Materials and Processes*. New York: McGraw-Hill, 1975.
68. Reiser, A. *Photoreactive Polymers—The Science and Technology of Resists*, New York: John Wiley and Sons, 1989.
69. Wallraff, G. M., and W. D. Hinsberg. "Lithographic Imaging Techniques for the Formation of Nanoscopic Features." *Chem. Rev.* 99 (1999): 1801.
70. McCoy, M. "Cleaner Chemistry for Cleaner Chips." *Chem. Eng. News* 79 (2001): 10.
71. Zeiger, D. H., T. M. Wolf, and G. N. Taylor. "Compressed Fluid Technology—Applications to Rie Developed Resists." *AIChE J.* 33 (1987): 1585.
72. Reichmanis, E., G. Smolinsky, and C. W. J. Wilkins. "Approaches to Resists for 2-Level Rie Pattern Transfer Applications." *Solid State Technol.* 28 (1985): 130.
73. Gallagher-Wetmore, P., G. M. Wallraff, and R. D. Allen. "Supercritical Fluid Processing: A New Dry Technique for Photoresist Developing." *Proc. SPIE* 2438 (1995): 694.
74. Ober, C. K., A. H. Gabor, P. Gallagher-Wetmore, and R. D. Allen. "Imaging Polymers with Supecritical Carbon Dioxide." *Adv. Mater.* 9 (1997): 1039.
75. DeSimone, J. M. *Proc. ACS Polym. Mater. Sci.* 79 (1998): 290.
76. Allen, R. D., K. J. Chen Rex, and P. M. Gallagher-Wetmore. "Performance Properties of Nearmonodisperse Novolak Resins." *Proc. SPIE* 2483 (1995): 250.
77. Gallagher-Wetmore, P., C. K. Ober, A. H. Gabor, and R. D. Allen. "Supercritical Fluid Processing: Opportunities for New Resist Materials and Processes." *Proc. SPIE* 2725 (1996): 289.
78. Allen, R. D. and G. M. Wallraff. IBM Corp., Process for generating negative tone resist images utilizing carbon dioxide critical fluid. U.S. Patent 5,665,527, 1997.
79. Sundararajan, N., S. Valiyaveettil, K. Ogino, X. Zhou, J. Wang, S. Yang, and C. K. Ober. "Block Copolymers as Supercritical CO_2 Developable Photoresists." *Proc. ACS Div. Polym. Mater. Sci. Eng.* 79 (1998): 130; Sundararajan, N., S. Yang, K. Ogino, S. Valiyaveettil, J.-G. Wang, X. Zhou, C. K. Ober, S. K. Obendorf, and R. D. Allen. "Supercritical CO_2 Processing for Submicron Imaging of Fluoropolymers." *Chem. Mater.* 1 (2000): 41.

80. Flowers, D., E. N. Hoggan, R. Carbonell, and J. M. DeSimone. "Designing Photoresist Systems for CO_2 Based Lithography." *Proc. SPIE* 4690 (2002): 419; Hoggan, E. N., K. Wang, D. Flowers, J. M. DeSimone, and R. Carbonell. "'Dry' Lithography Using Liquid and Supercritical Carbon Dioxide Based Chemistries and Processes." *IEEE Trans. Semicond. Manuf.* 17 (2004): 510.
81. Yang, S., J. Wang, K. Ogino, S. Valiyaveettil, and C. K. Ober. "Low-Surface-Energy Fluoromethacrylate Block Copolymers with Patternable Elements." *Chem. Mater.* 12 (2000): 33.
82. Bae, Y. C., K. Douki, T. Yu, J. Dai, D. Schmaljohann, H. Koerner, and C. K. Ober. "Tailoring Transparency of Imageable Fluoropolymers at 157 nm by Incorporation of Hexafluoroisopropyl Alcohol to Photoresist Backbones." *Chem. Mater.* 14 (2002): 1306.
83. Pham, V. Q., R. J. Ferris, A. Hamad, and C. K. Ober. "Positive-Tone Photoresist Process for Supercritical Carbon Dioxide Development." *Chem. Mater.* 15 (2003): 4893.
84. Millet, C., A. Danel, M. Ndour, and F. Tardif. "Compatibility of Supercritical CO_2-Based Stripping with Porous Ultra Low-κ Materials and Copper." *Solid State Phenom.* 92 (2003): 113.
85. Lebowitz, J., and L. E. Faulk. Residue removal by CO_2 water rinse in conjunction with post metal etch plasma strip. US Patent 63289095 B1, 2001.
86. Shieh, Y., J. Su, G. Manivannan, P. Lee, S. Sawan, and D. Spall. "Interaction of Supercritical Carbon Dioxide with Polymers: I. Crystalline Polymers." *J. Appl. Polym. Sci.* 59 (1996): 695.
87. Lambert, S. M., and M. E. Paulaitis. "Crystallization of Poly(ethylene terephthalate) Induced by Carbon Dioxide Sorption at Elevated Pressures." *J. Supercrit. Fluids* 4 (1991): 15.
88. Berens, A. R., G. S. Huvard, R. W. Korsmeyer, and F. W. Kunig. "Application of Compressed Carbon Dioxide in the Incorporation of Additives into Polymers." *J. Appl. Polym. Sci.* 46 (1992): 231.
89. Han, K. H., S. Y. Kim, and K. P. You. "The Study of Photoresist Removal by Using the Supercritical Carbon Dioxide." In *Conference Proceedings, 1st International Symposium on Supercritical Fluid Technology for Energy and Environment Applications—Super Green*, Suwon, Korea, November 3–6, 340, 2002.
90. Biberger, M. A., P. Schilling, D. Frye, and E. Mills. "Photoresist and Photoresist Residue Removal with Supercritical CO_2—A Novel Approach to Cleaning Wafers." *Semicond. FabTech.* 12 (2000): 239.
91. Barton, J. C. Apparatus and method for providing pulsed fluids. US Patent 6,085,762, 2000.
92. See www.scrub.lanl.gov website for additional information on the SCORR process.
93. Frazer, L. "SCORR One for the Environment." *Environ. Health Perspect.* 109 (2001): A383.
94. Rubin, J. B., L. B. Davenall, J. Barton, C. M. V. Taylor, and K. Tiefert. In *IEEE/CPMT Int. Electron. Manuf. Technol. Symp. 23rd*, New York, 308, 2001.
95. Papathomas, K. I., and A. C. Bhatt. "Debonding of Photoresist by Organic Solvents." *J. Appl. Polym. Sci.* 59 (1996): 2029.
96. Beyer, K. H. Jr., W. F. Bergfeld, W. D. Berndt, W. H. Carlton, D. K. Hoffman, A. L. Schroeter, and R. C. Shank. "Final Report on the Safety Assessment of Propylene Carbonate." *J. Am. Coll. Toxicol.* 6 (1987): 23.
97. Page, S. H., D. E. Raynie, S. R. Goates, M. L. Lee, D. J. Dixon, and K. P. Johnston. "Predictability and Effect of Phase Behavior of CO_2/Propylene Carbonate in Supercritical Fluid Chromatography." *J. Microcol. Sep.* 3 (1991): 355.
98. Rubin, J. B., L. D. Sivils, and A. A. Busnaina. "Precision Cleaning of Semiconductor Surfaces Using Supercritical Fluids." In *Proceedings of the Contamination Free Manufacturing Symposium*, San Francisco, July 12, 1, 1999.
99. Tolley, W. K., R. M. Izatt, and J. L. Oscarson. "Titanium Tetrachloride–Supercritical Carbon-Dioxide Interaction: a Solvent-Extraction and Thermodynamics STUDY." *Metall. Trans. B* 23 (1992): 65; Tolley, W. K., R. M. Izatt, J. L. Oscarson, R. L. Rowley, and N. F. Giles. "Comparison of the Thermodynamics of Mixing of Titanium Tetrachloride and Tin Tetrachloride with Supercritical Carbon Dioxide." *Sep. Sci. Technol.* 28 (1993): 615.
100. Giles, N. F., J. L. Oscarson, R. L. Rowley, W. K. Tolley, and R. M. Izatt. "Thermodynamic Properties of Mixing for $SnCl_4$ Dissolved in Supercritical CO_2: A Combined Experimental and Molecular-Dynamics Study." *Fluid Phase Equilib.* 73 (1992): 267.

101. Zhang, X., J. Q. Pham, P. J. Wolf, H. J. Martinez, P. F. Green, and K. P. Johnston. "Water-in-Carbon Dioxide Microemulsions for Removing Post-etch Residues from Patterned Porous Low-κ Dielectrics." *J. Vac. Sci. Technol. B* 21 (2003): 2590.
102. Muscat, A. J. "Supercritical Fluid Processing in Microelectronics Manufacturing." *Business Briefing: Global Semiconductor Manufacturing Technology*, 2003.
103. Wagner, M., J. DeYoung, S. Gross, Z. Hatcher, and C. Ma. In *ECS*, Pennington, NJ, 2003.
104. Myneni, S., and D. W. Hess. "Post-plasma Etch Residue Removal Using CO_2-Based Fluids." *J. Electrochem. Soc.* 150 (2003): G744.
105. DeGendt, S., D. Knotter, M. Heyns, M. Meuris, and P. Mertens. Method for removing organic contaminants from a semiconductor surface. US Patent 6551409, 2002.
106. DeYoung, J. P., J. B. McClain, and S. M. Gross. Methods for removing particles from microelectronic structures. US Patent Appl. 20020112746, 2002.
107. Korzenski, M. B., C. Xu, and T. H. Baum. "Supercritical Carbon Dioxide: The Next Generation Solvent for Semiconductor Wafer Cleaning Technology." In *Conference Proceedings, 6th Int. Symp. Supercrit. Fluids*, Versailles, France, April. 28–30, 2049, 2003.
108. Martinez, H. J., T. Jacobs, J. Wolf, and L. Rothman. "Supercritical Carbon Dioxide Processing of Porous Methylsilsesquiozane (PMSQ) Low-κ Dielectric Films." *Solid State Phenom.* 92 (2003): 293.
109. Wai, C. M., A. S. Gopalan, and H. K. Jacobs. "An Introduction to Separations and Processes Using Supercritical Carbon Dioxide." *ACS Symp. Ser.* 860 (2003): 2.
110. Laintz, K. E., C. M. Wai, C. R. Yonker, and R. D. Smith. "Extraction of Metal-Ions from Liquid and Solid Materials by Supercritical Carbon Dioxide." *Anal. Chem.* 64 (1992): 2875.
111. Glennon, J. D., J. Treacy, A. M. O'Keeffe, M. O'Connell, C. C. McSweeney, A. Walker, and S. J. Harris. "Extracting Gold in Supercritical CO_2: Fluorinated Molecular Baskets and Thiourea Ligands for Au." *ACS Symp. Ser.* 860 (2003): 67.
112. Kersch, C., M. J. E. Van Roosmalen, G. F. Woerlee, and G. J. Witkamp. "Extraction of Heavy Metals from Fly Ash and Sand with Ligands and Supercritical Carbon Dioxide." *Ind. Eng. Chem. Res.* 39 (2000): 4670.
113. Peters, L. "Supercritical CO_2 Finds Niche with Ultra Low-κ Materials." *Semicond. Int.* 8 (2003): 40.
114. Xu, C., D. W. Minsek, J. F. Roeder., M. B. Korzenski, and T. H. Baum. Supercritical cleaning of semiconductor substrates. US Patent Appl. 20030125225, 2003.
115. Wang, C. W., R. T. Chang, W. K. Lin, R. D. Lin, M. T. Liang, J. F. Yang, and J. B. Wang. "Supercritical CO_2 Fluid for Chip Resistor Cleaning." *J. Electrochem. Soc.* 146 (1999): 3485.
116. Maguire, J. F. *Proceedings of the Technical Program- National Electronic Packaging and Production Conference*, Des Plaines, IL. Vol. 2, 843, 1995.
117. Melton, C., and H. Fuerhaupter. "Lead Free Tin Surface Finish for PCB Assembly." *Circuit World* 23 (1997): 30.
118. Silaimani, S. M., M. Pushpavanam, and K. C. Narasinham. "Performance Characteristics of Electrochemically Prepared Tin Fluoborate." *Plat. Surf. Finish.* 83 (1996): 48.
119. Tijburg, R. "Advances in Etching of Semiconductor Devices." *Phys. Technol.* 7 (1976): 202.
120. DeYoung, J. P., S. M. Gross, M. I. Wagner, and J. B. McClain. Methods and compositions for etch cleaning microelectronic substrates in carbon Dioxide. U.S. Patent 6669785, 2003.
121. Jones, C. A. III., D. Yang, E. A. Irene, S. M. Gross, M. Wagner, J. DeYoung, and J. M. DeSimone. "HF Etchant Solutions in Supercritical Carbon Dioxide for 'Dry' Etch Processing of Microelectronic Devices." *Chem. Mater.* 15 (2003): 2867.
122. Xie, B., G. Montano-Miranda, C. C. Finstad, and A. J. Muscat. "Native Oxide Removal from SiGe Using Mixtures of HF and Water Delivered by Aqueous, Gas, and Supercritical CO_2 Processes." *Mater. Sci. Semicond. Process.* 8 (2005): 231.
123. Maboudian, R., and R. Howe. "Critical Review: Adhesion in Surface Micromechanical Structures." *J. Vac. Sci. Technol. B* 15 (1997): 1.
124. Jafri, I., H. Busta, and S. Walsh. "MEMS Reliability for Critical and Space Applications." In *SPIE Conference on MEMS Reliability for Critical and Space Applications*, 3880, Santa Clara, CA, 51, 1999.

125. Dyek, C. W., J. H. Smith, S. L. Miller, E. M. Russick, and C. L. J. Adkins. "Supercritical Carbon dioxide Solvent Extraction from Surface Micromachined Micromechanical Structures." In *SPIE Micromachining and Fabrication*, 1996.
126. Wallace, R. M., and M. A. Douglas. Method of cleaning and treating a semiconductor device including a micromechanical device. U.S. Patent 6024801, 1996.
127. Ho, C. M., and Y. C. Tai. "Review: MEMS and Its Applications for Flow Control." *J. Fluids Eng. - Trans. ASME* 118 (1996): 437.
128. Muraka, S. P., R. J. Gutmann, D. J. Duquette, and J. M. Steigerwald. Systems for performing chemical mechanical planarization and process for conducting same. U.S. Patent 5637185, 1997.
129. Steigerwald, J. M., S. P. Muraka, and R. J. Gumann. *Chemical Mechanical Planarization of Microelectronic Materials*. New York: Wiley, 1997.
130. McClain, J. B. and J. M. DeSimone. Methods, apparatus and slurries for chemical mechanical planarization. U.S. Patent 6743078, 2003.
131. Bessel, C. A., G. M. Denison, J. M. DeSimone, J. DeYoung, S. M. Gross, C. K. Schauer, and P. M. Visintin. "Etchant Solutions for the Removal of Cu(0) in a Supercritical CO_2-Based "Dry" Chemical Mechanical Planarization Process for Device Fabrication." *J. Am. Chem. Soc.* 125 (2003): 4980.
132. Douglas, M. A. and A. C. Templeton. Method for removing inorganic contamination by chemical derivatization and extraction, Patent 5,868,856, 1999; Douglas, M. A. and A. C. Templeton. Method of removing inorganic contamination by chemical alteration and extraction in a supercritical fluid media. Patent 5,868,862, 1999.
133. Xie, B., C. C. Finstad, and A. J. Muscat. "Removal of Copper from Silicon Surfaces Using Hexafluoroacetylacetone (hfacH) Dissolved in Supercritical Carbon Dioxide." *Chem. Mater.* 17 (2005): 1753.
134. Tanaka, T., M. Morigami, H. Oizumi, T. Soga, T. Ogawa, and F. Murai. "Prevention of Resist Pattern Collapse by Resist Heating during Rinsing." *J. Electrochem. Soc.* 141 (1994): L169.
135. Shibata, T., T. Ishii, H. Nozawa, and T. Tamamura. "High-Aspect-Ratio Nanometer-Pattern Fabrication Using Fullerene-Incorporated Nanocomposite Resists for Dry-Etching Application." *Jpn. J. Appl. Phys.* 36 (1997): 7642.
136. Tanaka, T., M. Morigami, H. Oizumi, T. Ogawa, and S. Uchino. "Prevention of Resist Pattern Collapse by Flood Exposure during Rinse Process." *Jpn. J. Appl. Phys.* 33 (1994): L1803.
137. Yamashita, Y. "Sub0.1 μm Patterning with High Aspect Ratio of 5 Achieved by Preventing Pattern Collapse." *Jpn. J. Appl. Phys.* 35 (1996): 2385.
138. Tanaka, T., M. Morigami, H. Oizumi, and T. Ogawa. "Freeze-Drying Process to Avoid Resist Pattern Collapse." *Jpn. J. Appl. Phys.* 32 (1993): 5813.
139. Canham, L. T., A. G. Cullis, C. Pickering, O. D. Dosser, T. I. Cox, and T. P. Lynch. "Luminescent Anodized Silicon Aerocrystal Networks Prepared by Supercritical Drying." *Nature* 368 (1994): 133.
140. Namatsu, H., K. Yamazaki, and K. Kurihara. "Supercritical Drying for Nanostructure Fabrication without Pattern Collapse." *Microelectron. Eng.* 46 (1999): 129.
141. Namatsu, H., K. Yamazaki, and K. Kurihara. "Supercritical Resist Dryer." *J. Vac. Sci. Technol. B* 18 (2000): 780.
142. Namatsu, H. "Supercritical Drying for Water Rinsed Resist Systems." *J. Vac. Sci. Technol. B* 18 (2000): 3308.
143. Goldfarb, D. L., J. J. dePablo, P. F. Nealey, J. P. Simons, W. M. Moreau, and M. Angelopoulos. "Aqueous-Based Photoresist Drying Using Supercritical Carbon Dioxide to Prevent Pattern Collapse." *J. Vac Sci. Technol. B* 18 (2000): 3313.
144. Xie, B., and A. J. Muscat. "Water Removal and Repair of Porous Ultra Low-κ Films Using Supercritical CO_2." In *Electrochemical Society Proceedings*, edited by J. Ruzyllo, T. Hattori, R. L. Opila, and R. E. Novak, 279. NJ: Pennington, 2003.
145. Xie, B., and A. J. Muscat. "Condensation of Silanol Groups in Porous Methylsilesquioxane Films Using Supercritical CO_2 and Alcohol Cosolvents." *IEEE Trans. Semicond. Manuf.* 17 (2004): 544.
146. Xie, B., and A. J. Muscat. "Silylation of Porous Methylsilsesquioxane Films in Supercritical Carbon Dioxide." *Microelectron. Eng.* 76 (2004): 52.

147. Xie, B., and A. J. Muscat. "Repair of Porous Methylsilsesquioxane Films Using Supercritical Carbon Dioxide." In *Materials, Technology and Reliability for Advanced Interconnects and Low-k Dielectrics, Material Research Society Symposium Proceedings*, 812, 13, 2004.
148. High Pressure Supercritical Carbon Dioxide Drying/Cleaning System. GT Equipment Technologies, Inc., 472 Amherst St. Nashua, NH 03063.
149. Douglas M. A. and R. M. Wallace. Method for unsticking components of micro-mechanical devices. US Patent 5,482,564, 1996.
150. Russick, E. M., C. L. J. Adkins, and C. W. Dyck. "Supercritical Carbon Dioxide Extraction of Solvent from Micromachined Structures, Supercritical Fluids—Extraction and Pollution Prevention." *ACS Symp. Ser.* 670 (1997): 255.
151. Perrut, V., J. V. Clavier, S. Lazure, A. Danel, and C. Millet. "Equipment for the Cleaning of Silicon Wafers Using Supercritical Fluids." In *Conference Proceedings, 6th International Symposium on Supercritical Fluids*, Versailles, France, April 28–30, 2073, 2003.
152. DeSimone, J. M., J. P. DeYoung, and J. B. McClain. Method and apparatus for cleaning substrates using liquid carbon dioxide. US Patent Appl. 20030051741, 2003.
153. Braun, A. E. "Photostrip Faces 300 mm, Copper and Low-κ Convergence." *Semicond. Int.* 23 (2000): 78.
154. Pacific northwest pollution prevention resource center, Supercritical carbon-dioxide cleaning technology review, 1996, Available online: http://www.pprc.org/pprc/p2tech/-co2/co2intro.html
155. Supercritical fluids research home page, Los Alamos National Laboratory, Available online: http://scrub.lanl.gov/html/scf/technologies/research_scorr_nn.htm
156. Supercritical fluid applications for advanced integrated circuit and MEMs fabrication, Available online: http://www.scfluids.com/news.htm
157. Rothman, L. "Supercritical CO_2 Tools and Applications for Semiconductor Processing." *Annual Retreat*, August 21–22, 2003, Available online: http://www.gscn.net/event/images/Speech-6.PDF

7

Ion Implantation

7.1 Introduction 7-1

7.2 Ion Implant Physics and Materials Science 7-2
Overview • Electronic and Nuclear Stopping and Collision Cascades • Implant Statistics • Channeling • Defects • List of Symbols Used in This Section

7.3 Applications of Ion Implantation 7-32
Introduction • Buried Layers • Source/Drain Implants • Source/Drain Extension Implants • Ultra-Shallow Junction Formation for Source/Drain Extensions

7.4 Commercial Ion Implantation Equipment 7-40
Introduction • Implanter Architecture Drivers • Ion Sources and Extraction Optics • Mass Analysis Systems • Post-Analysis Acceleration and Deceleration • Beam Scanning System and Dose Control • Wafer Charging Control • End Station

7.5 Process Control in Ion Implantation 7-61
Introduction • Beam-Residual Gas Interactions Control of Implanted Dose • Charging and Gate Oxide Integrity • Wafer Cooling and Photoresist Mask Integrity • Removal of Heavily Implanted Photoresists • Implant Angle Integrity • Wafer Contamination • Metrology

References 7-80

Michael Ameen
Ivan Berry
Walter Class
Hans-Joachim Gossmann
Leonard Rubin
Axcelis Technologies, Inc.

7.1 Introduction

In the past 30 years, ion implantation has emerged from laboratory tool status into a production process that is used today in virtually all semiconductor device manufacture. The reasons for this growth in popularity reside in the flexibility with which a variety of dopants can be selectively introduced into the semiconductor substrate, and the spatial and depth precision with which these dopants can be positioned. In today's nanotechnology world, great importance is placed on atom by atom construction, and on self-regulated assembly. These important attributes were achieved more than 20 years ago through the use of self-aligned ion implantation in the construction of complimentary metal oxide semiconductor (CMOS) transistor devices. These same attributes assure the continued use of this process as transistor device features assume atomic dimensions. Ion implantation is complex, and consequently a large body of knowledge related to the use and control of the process has accumulated over the years. This chapter provides a richly referenced overview of the following important features needed for the reader to gain an understanding of its application and control:

- knowledge of the physics and the materials science associated with the impact and penetration of energetic ions into a solid such as silicon,

- an overview of the body of ion implantation applications knowledge associated with the manufacture of state-of-art silicon CMOS devices,
- equipment architectural elements common to most commercial ion implantation systems, and their purpose, and
- an overview of important factors needed to assure controlled reproducible implant recipe execution.

7.2 Ion Implant Physics and Materials Science

7.2.1 Overview

Ion implantation is a process whereby energetic ions impinge on a target, penetrating below the target surface and giving rise to a controlled, predictable, and ion distribution. Here, we will focus on Si technology; hence the target will be mostly Si, but also including Si-oxides, Si-nitrides, and silicides. Implanted ions are typically dopants, such as Boron, Phosphorus, Arsenic, Indium, and Antimony. However, the scaling of device features into the sub-100 nm regime has added species, such as Ge, N, and Xe to this list. Implantation energies cover a wide range from 0.2 keV to more than 3 MeV; doses range from 1×10^{11} cm^{-2} to more than 1×10^{16} cm^{-2}; incident angles cover normal incidence (a tilt angle of 0°) to 60°.

Low implant energies produce dopant distributions near the surface such as required for metal oxide semiconductor (MOS) source and drain regions or bipolar emitter regions, and high energies produce deeply implanted dopant profiles such as required for CMOS retrograde wells and buried layers. The provision of wafer cooling during implant allows the use of photoresist masks to laterally control the location of dopant regions, an important feature for the production of CMOS devices. Similarly, topographical features of the device, such as a gate stack may be used to effect implant masking, thereby allowing for the production of cost- and yield-effective, self-aligned, doping regions.

A quantitative understanding of the atomic interactions between these energetic ions and the target atoms and masking layers is required for the prediction of the depth and lateral doping profiles produced by the ion implantation process. Intricately linked to ion implantation is the production of damage in the target, which also needs to be accurately predicted. This damage can take the form of a non-equilibrium excess of vacant lattice sites (vacancies) and self-interstitial atoms (interstitials), vacancy clusters, interstitial clusters, dopant-interstitial and dopant-vacancy clusters, and locally amorphized regions of the crystalline silicon target. Iso-valent ions, such as Si, Ge, or Sn are sometimes implanted to intentionally take advantage of this collateral damage. The annealing of this damage and the electrical activation of the implanted dopants, requires that the implanted target receive a subsequent heat treatment. The as-implanted defect configurations evolve during post-implant thermal processing, giving rise to transient enhanced dopant diffusion (TED), and the formation of relatively stable dislocation arrays, which if present in active device regions can lead to degradation of electrical performance. An understanding of all these phenomena is therefore crucial to the design of the implant recipe and the post-implant thermal treatment. The following reviews the underlying physics and materials science that give rise to the as-implanted dopant and defect distributions. The interaction of the defects with the dopants during annealing is discussed in Section 7.3.5.

7.2.2 Electronic and Nuclear Stopping and Collision Cascades

Energetic ion stopping occurs through interactions between the projectile ion and the electrons of the target (electronic stopping) and through the atomic collisions (nuclear stopping) between the projectile and the target atoms. These interactions slow the energetic ion until it eventual comes to a halt. The statistical distribution of the stopping points for many ions constitutes a concentration vs. depth profile. The prediction of such doping profiles, including the phenomenon of "channeling" is a central issue that underlies the interest in the stopping process. Stopping is characterized by an energy loss, dE,

per unit length of travel, dx, and is described by

$$\frac{dE}{dx} = S_n(E) + S_e(E), \tag{7.1}$$

where $S_n(E)$ and $S_e(E)$ are the nuclear and electronic stopping powers, respectively, and E is the energy of the ion. Both stopping components are a function of the ion energy, atomic number, and mass, and also the target atomic number, mass, atomic density, and electron distribution. The ion range results from the apportionment of the incident projectile energy over these stopping components. Frequently, the "stopping cross-section" is used, which is defined as

$$\varepsilon = \frac{S}{N}, \tag{7.2}$$

where N is the atomic density of the target. The fundamentals of atomic stopping were investigated [1] by Bohr in 1948. Lindhard et al. [2,3] developed them into a comprehensive theory, commonly referred to as Lindhard–Scharff–Schiott (LSS).

Electronic stopping results in excited target electronic states that decay without producing atomic displacements. On the other hand, nuclear stopping results from momentum and energy transfer between the projectile and the target atoms, which may result in the generation of energetic knock-on target atoms that may then produce secondary, tertiary, and higher order knock-on displacements. The resulting damage chain, known as a collision cascade, leaves behind a trail of vacant lattice sites along with implanted atoms and displaced target atoms in substitutional and interstitial positions. The collision-induced defects are predominantly produced as vacancy-interstitial pairs known as Frenkel defects, but can interact on the time-scale of the collision cascade (10^{-11} s) to recombine or agglomerate into more complex defect configurations. In addition, because of the initial momentum of the impinging projectile ions, the centroid of the interstitial distribution lies deeper in the target than the vacancy distribution.

7.2.2.1 Electronic Stopping

Electronic stopping is similar to the drag on a particle in a viscous medium, and involves excitation of the electrons of both the target and the projectile. It may also involve electron capture or loss from the bound states of the moving projectile ion. These processes determine the charge state and effective electronic stopping power of the ion as it moves through the solid.

7.2.2.1.1 Random Stopping

In the absence of channeling the projectile ion will see a spatially homogeneous electron density. Two major regimes in energy can be discriminated, depending on whether or not the velocity of the projectile, v, is much smaller than $v_c = Z_1 v_0$, where Z_1 is the projectile atomic number, $v_0 = e^2/2\varepsilon_0\,\hbar = 2.187 \times 10^6$ m/s the Bohr velocity, e the elementary charge, ε_0 the vacuum permittivity, and $\hbar$ the Planck constant.

7.2.2.1.1.1 Bethe–Bloch Regime

If $v \gg v_c$ the ion is stripped of its bound electrons. Here stopping can be calculated using the Bethe–Bloch theory [4,5]. The electronic stopping power in this regime, S_{eB}, is given as

$$S_{eB} = \frac{Z_1^2 Z_2 N}{m_e v^2} \frac{e^4}{4\pi\varepsilon_0^2} \ln \frac{2m_e v^2}{I} \tag{7.3}$$

where $I \approx 10\, Z_2$ is the average excitation potential, Z_2 the atomic number of the target atoms, and m_e the electron mass. However, such stopping is rarely encountered in ion implantation as Table 7.1 indicates, which lists energies corresponding to v_c for some common ion implant species.

It is quite obvious that the Bethe–Bloch regime need to be taken into account only for high energy implants of species lighter than P.

TABLE 7.1 Energy Corresponding to v_c of Commonly Implanted Ions

Species	E (MeV)
B	6.8
P	173
As	2021
Sb	7735

7.2.2.1.1.2 Lindhard–Scharff–Schiott Regime

At low ion velocities ($v \ll v_c$), the projectile ions are essentially neutral and stopping is proportional to the ion velocity, [2,3,6] giving the relationship as follows:

$$S_{eL}(E) = k_L\sqrt{E}. \tag{7.4}$$

The proportionality constant is given by

$$k_L = \sqrt{32}a_0\hbar\frac{Z_1^{7/6}Z_2N}{\left[Z_1^{2/3}+Z_2^{2/3}\right]^{3/2}M_1^{1/2}}, \tag{7.5}$$

where a_0 is the Bohr radius,

$$a_0 = \frac{\hbar^2\varepsilon_0}{\pi m_e e^2} \approx 5.293\times10^{-11}\text{m}, \tag{7.6}$$

and M_1 the mass of the projectile atom. Beyond P it varies little with Z_1, i.e., the stopping power is essentially independent of species, as Figure 7.1 indicates.

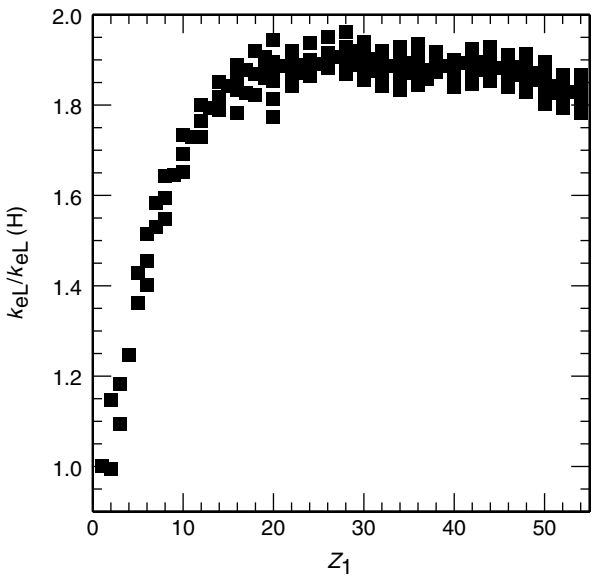


FIGURE 7.1 Values of k_L normalized to the value for H, for different ion species implanted into Si.

7.2.2.1.1.3 Interpolation Between Low Velocity and High Velocity Regimes

Biersack and Haggmark [7] proposed an interpolation scheme for a smooth transition between the LSS and the Bethe–Bloch regimes by setting

$$S_e = \frac{1}{S_{eL}^{-1} + S_{eB}^{-1}} \tag{7.7}$$

Figure 7.2 shows the resulting electronic stopping power for implantation of B, P, and As into Si.

7.2.2.1.1.4 Practical Considerations

The LSS and Bethe–Bloch theories yield analytical formulae but by necessity, make various approximations. For practical application they usually are modified [8–10]. Equation 7.4 is written as

$$S_{er}(E) = kE^r \tag{7.8}$$

Equation 7.7 is written as

$$S_e = \left[S_{er}^{-c} + S_{eB}^{-c}\right]^{-1/c}, \tag{7.9}$$

The former becomes Equation 7.4 if $k/k_L = 1$ and $r = 0.5$; the latter reverts to Equation 7.7 if $c = 1$. The various authors have derived different best fits of k/k_L, r, and c to experimental range data, such as $k/k_L = 1.5$, $r = 0.5$, and $c = 0.9$ by Simionescu et al. [8] and $r = 0.375$, and $c = 1.0$ by Ziegler et al. [10].

7.2.2.1.2 Channeling

The assumption of an uniform electron distribution in the target, and of an energy loss that is independent of the relative positions of the projectile ion and the target atoms ("non-local model" [11]) is no longer justified if the target is single-crystalline: the electron density in the center of a low-index channel is very low. Accurate description of electronic stopping requires a "local model" where

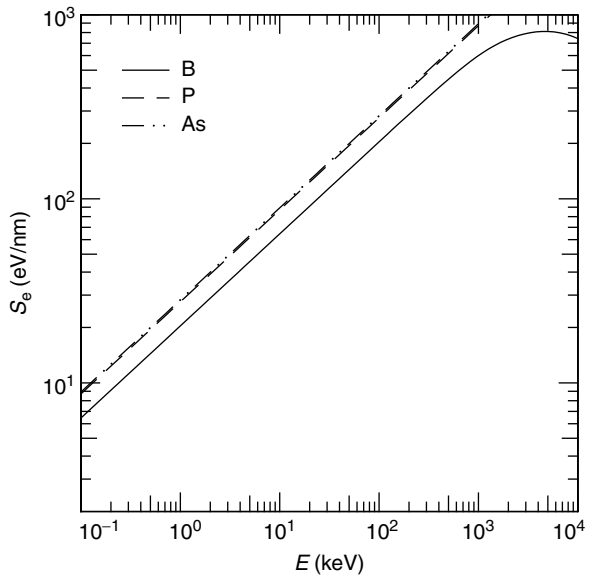


FIGURE 7.2 Electronic stopping power for implantation of B, P, and As into Si. The curves for P and As fall essentially on top of each other.

a position-dependent energy loss is explicitly taken into account, either via the impact parameters of the collisions [12,13] or using the electron concentration along the projectile's path [14]. Hobler et al. have reported [11] that the latter failed with realistic electron density distributions and that the non-local- and the local impact parameter-dependent model each fail to describe both random and channeling conditions. Therefore, they proposed a combination of both, i.e., the electronic energy loss has a non-local part given by

$$\Delta E_{\mathrm{e,nl}} = NS_{\mathrm{e}}\Delta R\left[x_{\mathrm{nl}} + x_{\mathrm{loc}}\left(1 + \frac{b_{\max}}{a}\right)\exp\left(-\frac{b_{\max}}{a}\right)\right] \tag{7.10}$$

and a local part is given by

$$\Delta E_{\mathrm{e,loc}} = x_{\mathrm{loc}}\frac{S_{\mathrm{e}}}{2\pi a^2}\exp\left(-\frac{b}{a}\right) \tag{7.11}$$

Here, b is the impact parameter and $b_{\max}$ its maximum value, x_{loc} and x_{nl} are the local and non-local fractions, respectively, and a is the screening length. To make random electronic stopping independent of those parameters, x_{loc} and x_{nl} are constrained by

$$x_{\mathrm{loc}} + x_{\mathrm{nl}} = 1. \tag{7.12}$$

The non-local fraction is found [11] to be energy-dependent as

$$x_{\mathrm{nl}} = y_{\mathrm{nl}}E^q \tag{7.13}$$

The electronic stopping power S_{e} is given by Equation 7.5 through Equation 7.8 with $r=0.5$ and $k/k_L=1.5$. The best fit of the three free parameters to implantation data of B into Si yields [11] $q=0.12$, $y_{\mathrm{nl}}=0.16\ \mathrm{eV}^{-0.12}$, and $a=2a_{\mathrm{OR}}$, where a_{OR} is the screening length of the interatomic potential [10]. Nevertheless, despite all the effort expended, the simulation of electronic stopping in single-crystalline targets still is not quite satisfactory and continues to be an actively researched topic [15].

7.2.2.2 Nuclear Stopping

Nuclear stopping is due to the energy transfer between a projectile and a stationary target ion during the collision. This scattering event is schematically shown in Figure 7.3. A projectile ion of atomic number Z_1 and atomic mass M_1 approaches with velocity v_1 the target atom of atomic number Z_2 and atomic mass M_2 that is originally stationary. The path of the projectile ion is offset from the center of the target atom by the impact parameter, b. The interactions between projectile and target atom scatters the former through a scattering angle Θ and it moves on with velocity v_1' the target atom acquires a velocity v_2'.

The energy transfer is given by

$$T = \frac{4M_1M_2}{(M_1+M_2)^2}E_1\sin^2\frac{\Theta}{2}, \tag{7.14}$$

where E_1 is the energy of the projectile. The scattering angle, Θ, depends on the interaction potential, $V(r)$. An impact parameter of $b=0$ leads to a maximum scattering angle and a maximum energy transfer, $T_{\max}$. Given $V(r)$ the differential scattering cross-section $\mathrm{d}\sigma/\mathrm{d}\Omega$, with Ω the solid angle, may be calculated and the nuclear stopping power then is given by

$$S_{\mathrm{n}} = N\int_0^{\Theta(T_{\max})} T\,\mathrm{d}\sigma. \tag{7.15}$$

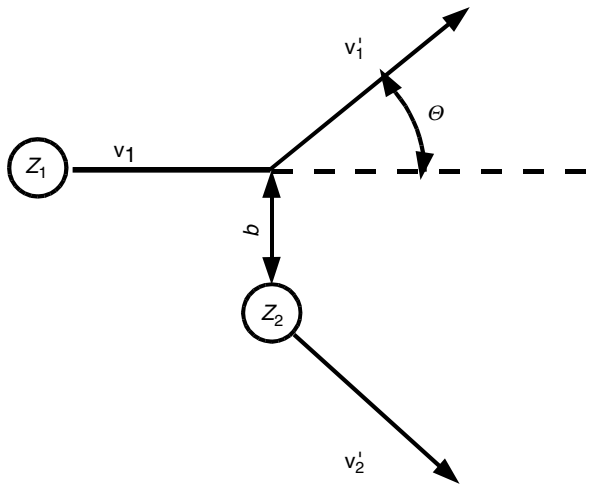


FIGURE 7.3 Scattering of an incident ion, (Z_1, M_1, v_1), and a target atom, (Z_2, M_2, $v_2=0$). After the collision the velocities are v_1' and v_2', respectively.

While conceptually simple the accuracy of S_n is only as good as the potential V is known. Typically it is expressed in terms of a screening function, $f_s(r)$, such that

$$V(r) = \frac{Z_1 Z_2 e^2}{4\pi\varepsilon_0 r} f_s(r). \tag{7.16}$$

The various approximations for the interactions between the projectile and the target atom, including their respective electron clouds, have been attempted. The most accurate involve Hartree–Fock models of the atom, but, since they are very computer-intensive, are only used to validate simpler approaches. Lindhard, Scharff, and Schiott [3] used the Thomas–Fermi approximation of the interaction potential V, i.e., a screening function of the form

$$f_s(r) = e^{-r/a} \tag{7.17a}$$

where

$$a = \frac{0.8853 a_0}{\left(Z_1^{2/3} + Z_2^{2/3}\right)^{1/2}}, \tag{7.17b}$$

to obtain analytical expressions for S_n. Ziegler, Biersack, and Littmark (ZBL) [10] used the simplified quantum-mechanical approach originally suggested by Gombas [16] to approximate the interaction potentials, resulting in the widely used dataset of the "ZBL potentials." By further fitting a sum of exponentials of the form in Equation 7.17a they obtained an analytical "Universal Screening Function," valid for all projectile-target combinations, and given by

$$f_U = \sum_{i=1}^{4} f_{U,i} e^{-a_{U,i}\frac{r}{a_U}} \tag{7.18a}$$

with

$$a_U = \frac{0.8854 a_0}{\left(Z_1^{0.23} + Z_2^{0.23}\right)^{1/2}}, \tag{7.18b}$$

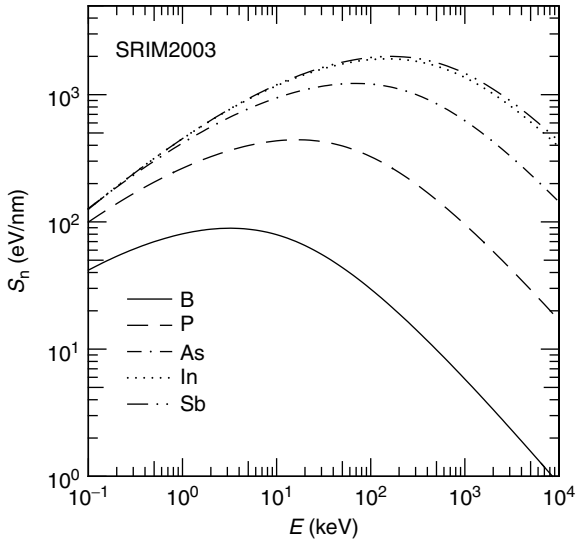


FIGURE 7.4 Nuclear stopping power for implantation of B, P, As, In, and Sb into Si.

$f_{U,i}$ and $a_{U,i}$ are given by the vectors

$$f_{U,i} = (0.1818, 0.5099, 0.2802, 0.02817) \tag{7.18c}$$

and

$$a_{U,i} = (3.2, 0.9423, 0.4028, 0.216), \tag{7.18d}$$

respectively. By integrating Equation 7.15 with the reduced variables suggested by Lindhard, Scharff, and Schiott [3] an universal nuclear stopping power is then obtained [10].

Figure 7.4 shows the nuclear stopping power for some common dopants implanted into Si, as calculated by the stopping and range of ions in solids (SRIM)-2003 [17].

7.2.3 Implant Statistics

7.2.3.1 Gaussian Profiles

The range, R, of an ion may simply be obtained by integrating Equation 7.1:

$$E_1 = \int_0^R S\,\mathrm{d}x. \tag{7.19}$$

However, this is the distance traveled by an ion along its track and hence of limited value. Far more important is the projected range, R_p, the distance that the projectile ions on average penetrate into the target; Figure 7.5 shows the relationships between R and R_p.

Lindhard, Scharff, and Schiott [3] applied statistical concepts and the electronic and the nuclear stopping discussed in the preceding section to arrive at an analytical expression for the mean range, R, and its standard deviation, the straggle σ. Their derivation also allowed for the determination of R_p and σ_p. Since [2] $\sigma_p \approx \sigma$ we will drop the subscript "p" from the straggle in the remainder of this text.

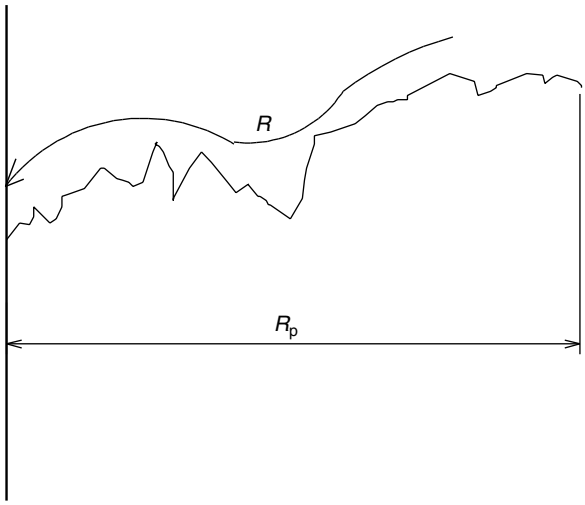


FIGURE 7.5 Schematic illustrating the relationship between R and R_p.

Projected range and straggle are the first and second moment of the implanted ion distribution,

$$R_p = \frac{1}{n}\sum x_i \tag{7.20}$$

and

$$\sigma^2 = \frac{1}{n}\sum (x_i - R_p)^2, \tag{7.21}$$

where the summation extends over all ions, of which there are a total of n, and x_i is the depth at which they come to rest.

The general shape of the implanted ion distribution is then Gaussian, i.e., the volume concentration of implanted ions A is given by

$$N_A = \frac{\Phi}{\sqrt{2\pi\sigma}} e^{-\frac{(x-R_p^2)}{2\sigma^2}}, \tag{7.22}$$

where Φ is the implanted dose per unit area. The peak concentration is

$$N_{A,max} = \frac{\Phi}{\sqrt{2\pi\sigma}} \approx \frac{0.4}{\sigma} \tag{7.23}$$

We note that Equation 7.22 only applies to amorphous targets.

7.2.3.2 The Pearson-IV Distribution

As we will see below, crystalline targets do not lead to Gaussian profiles, especially if the beam is incident along a major crystallographic axis ("channeling"). Even if this is not the case, the significant deviations of the dopant distribution from a Gaussian shape occur, specifically, the distribution is asymmetric with an extensive, non-Gaussian tail. Such shapes are far better described by a Pearson-IV distribution [18] and the first four moments,

$$\mu_1 = R_p, \tag{7.24a}$$
$$\mu_2 = \sigma^2, \tag{7.24b}$$
$$\mu_3 = \frac{1}{n}\sum (x_i - \mu_1)^3, \tag{7.24c}$$
$$\mu_4 = \frac{1}{n}\sum (x_i - \mu_1)^4 \tag{7.24d}$$

Usually, the third and the fourth moment are written in an alternative form, as "Skewness,"

$$\gamma = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}, \tag{7.25a}$$

and "Kurtosis,"

$$\beta = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}, \tag{7.25b}$$

respectively. Both are dimensionless numbers.

The concentration of a dopant A is then given by

$$N_A = C\left[1 + \left(\frac{x-\lambda}{a}\right)^2\right]^{-m} \exp\left[-\nu \tan^{-1}\left(\frac{x-\lambda}{a}\right)\right], \tag{7.26}$$

where C is a normalization constant such that

$$\Phi = \int_0^\infty N_A(x)\mathrm{d}x. \tag{7.27}$$

Using the abbreviation $r=2(m-1)$, the moments of a Pearson-IV distribution can be calculated readily from the parameters m, ν, a, and λ via [18]

$$R_p = \lambda - \frac{a\nu}{r}, \tag{7.28a}$$
$$\sigma^2 = \frac{a^2}{r^2(r-1)}(r^2 + \nu^2), \tag{7.28b}$$
$$\gamma = \frac{-4\nu}{r-2}\sqrt{\frac{r-1}{r^2+\nu^2}}, \tag{7.28c}$$
$$\beta = \frac{3(r-1)\left[(r+6)(r^2+\nu^2) - 8\nu^2\right]}{(r-2)(r-3)(r^2+\nu^2)}. \tag{7.28d}$$

In practice, it is the moments that are known and the Pearson-IV parameters that are sought [18]:

$$r = 2(m-1) = \frac{6(\beta - \gamma^2 - 1)}{2\beta - 3\gamma^2 - 6} \tag{7.29a}$$
$$\nu = -\frac{r(r-2)\gamma}{\sqrt{16(r-1) - \gamma^2(r-2)^2}} \tag{7.29b}$$
$$a = \frac{\sigma}{4}\sqrt{16(r-1) - \gamma^2(r-2)^2} \tag{7.29c}$$
$$\lambda = R_p - \frac{1}{4}(r-2)\gamma\sigma. \tag{7.29d}$$

while a Pearson-IV distribution describes implant statistics in amorphous material fairly well [19] in crystalline targets it has been found that ion channeling, described below, can result in a bimodal profile, the fitting of which requires a Dual Pearson-IV model.

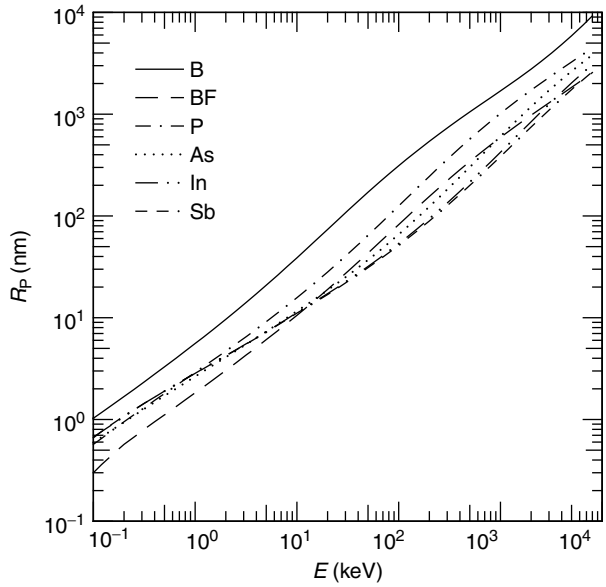


FIGURE 7.6 Projected range as a function of energy for common dopants implanted into amorphous Si (IMSIL-2003 simulation).

Figure 7.6 through Figure 7.10 show the first four moments as well as the lateral straggle for B, P, As, In, and Sb implanted into amorphous Si, as a function of energy. Table 7.2 through Table 7.6 give the numerical values and the fraction that is backscattered ("Doseloss"). A Monte Carlo (MC) simulator, IMSIL [20] (see Section 7.2.3.3) was used to obtain this data.

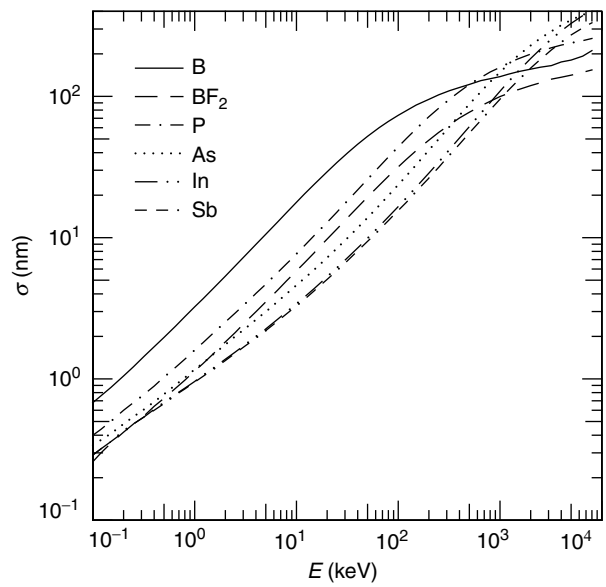


FIGURE 7.7 Straggle as a function of energy for common dopants implanted into amorphous Si (IMSIL-2003 simulation).

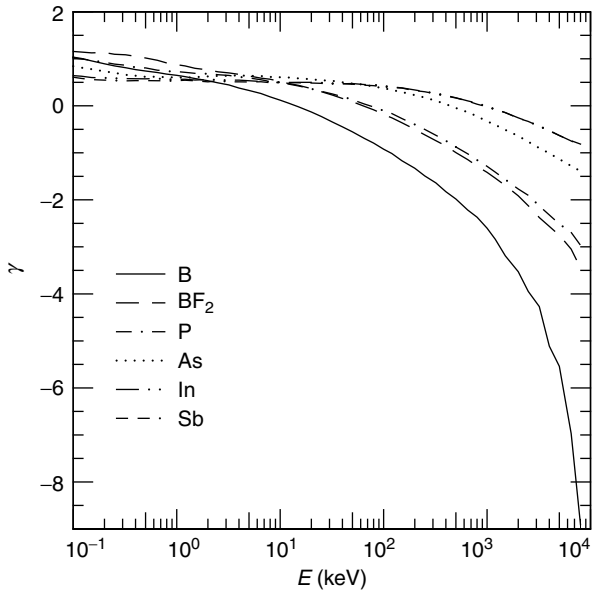


FIGURE 7.8 Skewness as a function of energy for common dopants implanted into amorphous Si (IMSIL-2003 simulation).

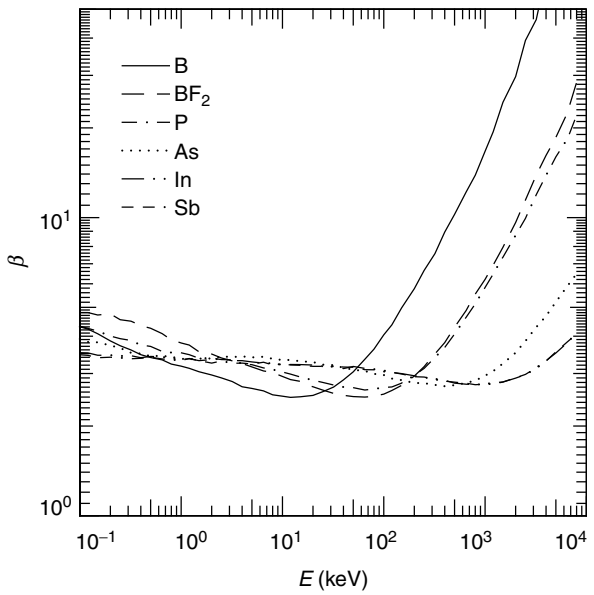


FIGURE 7.9 Kurtosis as a function of energy for common dopants implanted into amorphous Si (IMSIL-2003 simulation).

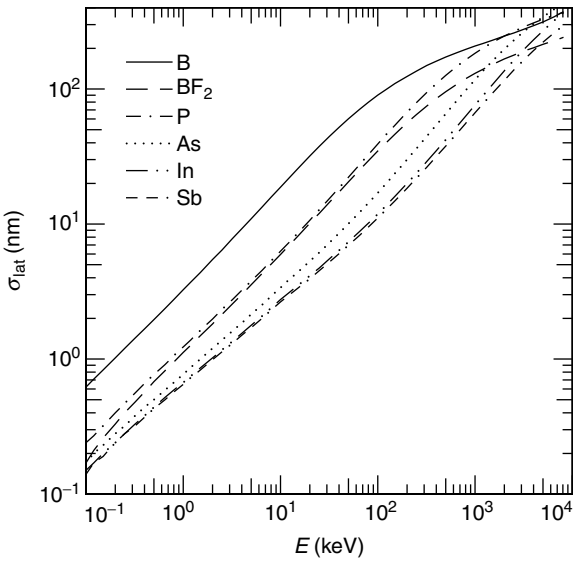


FIGURE 7.10 Lateral straggle as a function of energy for common dopants implanted into amorphous Si (IMSIL-2003 simulation).

7.2.3.3 Monte Carlo Approaches

A Pearson-IV distribution represents the dopant distribution quite well in amorphous targets. Figure 7.11 illustrates this by comparing a MC simulation [20] of 20 keV B $2\times10^{14}\,\mathrm{cm}^{-2}$ into amorphous Si with the Pearson-IV distribution using the moments extracted from the simulated depth profile.

When the target is crystalline channeling cannot be avoided even if implant angles are chosen to make the target appear amorphous ("random incidence"; see Section on "Channeling") and the distribution becomes deeper and bimodal. Figure 7.12 illustrates a kink that is seen at a depth of around 150 nm and the agreement between the Pearson-IV distribution and the concentration profile degrades. Even larger deviations occur for a beam perfectly aligned along a major crystallographic axis, as illustrated in Figure 7.13.

Therefore, process simulations that use predetermined moments to calculate dopant distributions analytically need to be viewed with caution, in particular if the simulation is 2D and even more so if the implant takes place into a non-planar structure. While moment-based approaches are undoubtedly very fast and useful for quick surveys, accurate, and predictive depth profiles will only be obtained from the simulation of the actual physical processes.

There are two possible approaches; the most accurate, but also the most time-consuming in terms of computer resources, are molecular dynamics (MD) simulations. Less complex and significantly faster in execution are MC techniques. They are based on the binary-collision approximation, tracking the motion of implanted ions deterministically. Contrary to MD, only the interactions with the nearest atoms are taken into account. This assumption is valid as long as the kinetic energy of the implanted ions is much higher than the potential energy arising from the interatomic pair potential of the ion and the target atoms.

Monte Carlo implant simulators are incorporated into commercial process simulation software but stand-alone programs are also available. The range of ions in solids (TRIM)/SRIM [17] is a 1D simulator that uses the ZBL potentials. Since it also assumes an amorphous target and is only available on a proprietary platform (Windows) its usefulness is somewhat limited. However it also produces range

TABLE 7.2 Implant Statistics of B Implanted into Amorphous Si

E (keV)	R_p (nm)	σ (nm)	γ	β	$R_{p,lat}$ (nm)	σ_{lat} (nm)	Dose-Loss (%)
0.10	1.03	0.68	1.034	4.305	0.16	0.62	9.3
0.12	1.17	0.76	1.006	4.230	0.17	0.71	9.2
0.15	1.37	0.87	0.956	4.077	0.19	0.83	8.9
0.20	1.68	1.05	0.893	3.858	0.22	1.02	8.8
0.25	1.97	1.22	0.848	3.719	0.26	1.20	8.6
0.32	2.36	1.45	0.811	3.598	0.29	1.43	8.2
0.40	2.79	1.69	0.769	3.471	0.34	1.67	8.0
0.50	3.29	1.97	0.740	3.395	0.38	1.96	7.6
0.65	4.01	2.37	0.709	3.332	0.47	2.37	7.3
0.80	4.70	2.76	0.673	3.227	0.55	2.76	7.3
1.00	5.58	3.24	0.649	3.175	0.64	3.26	6.9
1.20	6.42	3.68	0.620	3.123	0.73	3.73	6.8
1.50	7.65	4.33	0.580	3.042	0.86	4.38	6.4
2.00	9.65	5.34	0.534	2.965	1.08	5.46	5.9
2.50	11.58	6.31	0.489	2.900	1.26	6.42	5.6
3.20	14.22	7.58	0.443	2.818	1.59	7.80	5.1
4.00	17.19	8.96	0.378	2.709	1.88	9.22	4.8
5.00	20.89	10.58	0.327	2.674	2.29	10.94	4.3
6.50	26.30	12.86	0.262	2.624	2.95	13.43	3.7
8.00	31.73	15.06	0.191	2.545	3.48	15.81	3.3
10.00	38.76	17.80	0.120	2.517	4.26	18.75	2.8
12.00	45.94	20.40	0.057	2.496	5.21	21.59	2.5
15.00	56.54	23.93	−0.021	2.506	6.31	25.63	2.1
20.00	73.60	29.25	−0.131	2.533	8.34	31.92	1.6
25.00	90.64	34.10	−0.227	2.608	10.14	37.61	1.3
32.00	114.06	39.94	−0.345	2.702	13.00	44.76	1.0
40.00	139.76	45.88	−0.447	2.855	15.69	52.00	0.8
50.00	171.12	52.17	−0.554	3.038	19.15	60.25	0.6
65.00	215.83	60.17	−0.693	3.349	24.51	70.72	0.4
80.00	258.18	66.52	−0.793	3.620	29.40	79.77	0.3
100.00	311.38	73.49	−0.923	4.044	35.12	90.21	0.2
120.00	362.03	79.25	−1.015	4.365	41.48	98.44	0.2
150.00	432.58	86.91	−1.168	4.980	49.29	109.84	0.1
200.00	541.50	95.32	−1.327	5.775	61.88	123.86	0.1
250.00	640.53	102.63	−1.492	6.652	72.74	135.89	0.1
320.00	767.16	109.77	−1.639	7.571	87.45	149.52	0.0
400.00	900.10	115.67	−1.827	8.943	102.17	160.76	0.0
500.00	1052.20	121.52	−1.983	10.249	119.16	172.54	0.0
650.00	1259.98	129.45	−2.216	12.187	143.55	185.93	0.0
800.00	1452.92	133.36	−2.367	13.879	165.34	196.54	0.0
1000.00	1693.06	138.08	−2.603	16.690	193.24	208.32	0.0
1200.00	1920.61	143.47	−2.840	19.381	219.17	217.26	0.0
1500.00	2248.11	149.53	−3.190	24.377	255.80	229.42	0.0
2000.00	2774.06	155.94	−3.527	29.703	315.15	246.88	0.0
2500.00	3289.61	161.16	−3.947	39.288	374.53	260.18	0.0
3200.00	4009.80	165.28	−4.270	46.138	455.97	277.89	0.0
4000.00	4838.50	175.30	−5.112	65.427	550.81	294.78	0.0
5000.00	5894.49	180.69	−5.541	79.154	670.76	314.03	0.0
6500.00	7534.23	192.94	−6.970	123.287	857.05	342.53	0.0
8000.00	9251.28	211.03	−8.906	186.552	1053.30	371.86	0.0

TABLE 7.3 Implant Statistics of B Implanted as BF_2 into Amorphous Si

E (keV)	R_p (nm)	σ (nm)	γ	β	$R_{p,lat}$ (nm)	σ_{lat} (nm)	Dose-Loss (%)
0.10	0.30	0.26	1.155	4.805	0.09	0.17	0.9
0.12	0.36	0.30	1.144	4.803	0.09	0.21	2.4
0.15	0.44	0.35	1.125	4.713	0.10	0.26	4.7
0.20	0.57	0.42	1.133	4.752	0.11	0.33	7.4
0.25	0.67	0.48	1.096	4.546	0.12	0.40	8.4
0.32	0.81	0.55	1.086	4.493	0.13	0.48	9.2
0.40	0.95	0.64	1.052	4.341	0.15	0.57	9.4
0.50	1.11	0.73	1.018	4.254	0.16	0.68	9.6
0.65	1.34	0.86	0.969	4.082	0.18	0.82	9.2
0.80	1.56	0.98	0.912	3.889	0.21	0.95	9.2
1.00	1.83	1.14	0.876	3.797	0.23	1.12	8.8
1.20	2.08	1.28	0.848	3.737	0.26	1.27	8.6
1.50	2.45	1.49	0.798	3.549	0.29	1.49	8.4
2.00	3.03	1.83	0.763	3.487	0.36	1.82	8.0
2.50	3.58	2.13	0.723	3.381	0.41	2.14	7.9
3.20	4.33	2.56	0.700	3.307	0.49	2.56	7.7
4.00	5.11	2.99	0.668	3.241	0.57	3.01	7.4
5.00	6.09	3.51	0.634	3.157	0.67	3.55	7.2
6.50	7.47	4.24	0.589	3.060	0.82	4.32	6.8
8.00	8.81	4.95	0.565	3.022	0.96	5.04	6.5
10.00	10.56	5.82	0.518	2.933	1.17	5.95	6.2
12.00	12.29	6.65	0.475	2.863	1.31	6.84	5.7
15.00	14.82	7.85	0.436	2.823	1.64	8.10	5.2
20.00	18.97	9.77	0.367	2.707	2.11	10.12	4.6
25.00	23.09	11.53	0.300	2.640	2.58	12.00	4.1
32.00	28.82	13.92	0.228	2.573	3.12	14.54	3.5
40.00	35.25	16.44	0.157	2.534	3.91	17.26	3.1
50.00	43.24	19.44	0.086	2.510	4.80	20.56	2.6
65.00	55.20	23.46	−0.011	2.501	6.17	25.14	2.1
80.00	66.70	27.22	−0.085	2.514	7.38	29.41	1.8
100.00	82.13	31.72	−0.182	2.559	9.27	34.74	1.5
120.00	97.28	35.85	−0.256	2.631	11.11	39.55	1.2
150.00	119.69	41.25	−0.360	2.738	13.54	46.26	0.9
200.00	155.39	49.05	−0.500	2.942	17.74	56.23	0.7
250.00	189.75	55.69	−0.617	3.164	21.56	64.78	0.5
320.00	235.36	63.04	−0.741	3.494	26.93	74.91	0.4
400.00	284.74	70.35	−0.855	3.808	32.62	85.00	0.3
500.00	343.15	77.25	−0.996	4.291	39.32	95.49	0.2
650.00	423.74	85.77	−1.144	4.882	48.00	108.53	0.1
800.00	498.80	92.37	−1.282	5.543	56.79	119.19	0.1
1000.00	590.89	99.71	−1.416	6.205	67.25	130.69	0.1
1200.00	677.23	104.86	−1.538	6.928	77.48	139.89	0.1
1500.00	796.38	111.54	−1.695	7.951	90.78	151.72	0.0
2000.00	976.15	119.22	−1.914	9.628	111.11	167.24	0.0
2500.00	1140.21	125.08	−2.124	11.442	129.11	178.25	0.0
3200.00	1349.98	132.56	−2.377	13.886	153.63	191.62	0.0
4000.00	1572.67	136.80	−2.565	16.169	179.24	202.08	0.0
5000.00	1834.20	141.42	−2.754	18.589	208.38	214.26	0.0
6500.00	2205.66	148.02	−3.047	22.499	251.25	228.22	0.0
8000.00	2562.12	154.03	−3.446	28.415	291.05	240.58	0.0

TABLE 7.4 Implant Statistics of P Implanted into Amorphous Si

E (keV)	R_p (nm)	σ (nm)	γ	β	$R_{p,lat}$ (nm)	σ_{lat} (nm)	Dose-Loss (%)
0.10	0.57	0.40	1.020	4.322	0.12	0.24	0.6
0.12	0.65	0.44	0.996	4.251	0.13	0.27	0.6
0.15	0.77	0.50	0.967	4.164	0.14	0.32	0.6
0.20	0.95	0.60	0.924	4.006	0.16	0.40	0.6
0.25	1.11	0.68	0.888	3.926	0.18	0.47	0.5
0.32	1.31	0.78	0.862	3.871	0.20	0.56	0.4
0.40	1.54	0.89	0.819	3.739	0.22	0.66	0.4
0.50	1.80	1.03	0.806	3.699	0.25	0.77	0.4
0.65	2.15	1.21	0.758	3.557	0.29	0.91	0.4
0.80	2.49	1.38	0.737	3.526	0.33	1.06	0.4
1.00	2.90	1.60	0.729	3.520	0.38	1.23	0.3
1.20	3.29	1.80	0.712	3.471	0.42	1.39	0.3
1.50	3.85	2.09	0.690	3.376	0.48	1.63	0.3
2.00	4.69	2.53	0.681	3.378	0.58	1.98	0.3
2.50	5.50	2.93	0.662	3.342	0.67	2.33	0.3
3.20	6.55	3.46	0.646	3.335	0.78	2.75	0.3
4.00	7.73	4.03	0.616	3.231	0.91	3.23	0.3
5.00	9.14	4.69	0.594	3.211	1.06	3.78	0.3
6.50	11.20	5.65	0.555	3.132	1.29	4.56	0.2
8.00	13.16	6.54	0.517	3.095	1.54	5.33	0.2
10.00	15.73	7.68	0.490	3.034	1.84	6.29	0.2
12.00	18.26	8.79	0.464	2.987	2.13	7.20	0.2
15.00	21.97	10.38	0.422	2.921	2.53	8.54	0.2
20.00	28.12	12.82	0.361	2.847	3.26	10.67	0.1
25.00	34.17	15.19	0.309	2.800	3.89	12.75	0.1
32.00	42.70	18.38	0.252	2.744	4.93	15.52	0.1
40.00	52.36	21.81	0.194	2.702	6.07	18.51	0.1
50.00	64.51	25.90	0.131	2.690	7.38	22.23	0.1
65.00	82.78	31.78	0.038	2.642	9.39	27.58	0.1
80.00	101.16	37.19	−0.025	2.661	11.55	32.64	0.0
100.00	125.44	43.93	−0.105	2.693	14.24	39.16	0.0
120.00	149.68	50.33	−0.166	2.703	17.19	45.46	0.0
150.00	185.93	58.90	−0.271	2.786	21.34	53.89	0.0
200.00	244.89	71.92	−0.397	2.954	28.00	67.43	0.0
250.00	303.03	82.47	−0.503	3.121	34.43	79.42	0.0
320.00	381.60	95.61	−0.627	3.397	43.30	94.56	0.0
400.00	467.75	108.13	−0.744	3.668	53.48	109.59	0.0
500.00	570.06	121.47	−0.876	4.071	64.55	126.08	0.0
650.00	714.24	136.40	−1.021	4.632	81.43	147.26	0.0
800.00	848.60	148.75	−1.153	5.136	96.57	164.56	0.0
1000.00	1016.10	160.68	−1.287	5.832	116.29	183.91	0.0
1200.00	1173.53	171.29	−1.410	6.468	133.59	200.26	0.0
1500.00	1391.08	184.41	−1.567	7.369	157.59	220.84	0.0
2000.00	1721.89	199.15	−1.769	8.720	195.71	247.04	0.0
2500.00	2019.96	209.45	−1.903	9.942	231.46	266.57	0.0
3200.00	2402.51	221.05	−2.129	11.935	274.86	289.36	0.0
4000.00	2800.41	230.88	−2.322	13.848	318.73	308.98	0.0
5000.00	3258.51	239.01	−2.516	15.982	371.35	328.83	0.0
6500.00	3889.38	248.00	−2.691	18.393	442.03	350.38	0.0
8000.00	4471.89	257.05	−2.966	22.006	509.91	368.79	0.0

TABLE 7.5 Implant Statistics of As Implanted into Amorphous Si

E (keV)	R_p (nm)	σ (nm)	γ	β	$R_{p,lat}$ (nm)	σ_{lat} (nm)	Dose-Loss (%)
0.10	0.59	0.34	0.843	3.917	0.12	0.17	0.0
0.12	0.67	0.37	0.811	3.826	0.13	0.20	0.0
0.15	0.78	0.41	0.780	3.816	0.14	0.23	0.0
0.20	0.94	0.48	0.733	3.725	0.16	0.28	0.0
0.25	1.09	0.54	0.702	3.653	0.18	0.32	0.0
0.32	1.28	0.61	0.658	3.520	0.20	0.38	0.0
0.40	1.48	0.69	0.641	3.499	0.22	0.44	0.0
0.50	1.71	0.78	0.617	3.432	0.24	0.50	0.0
0.65	2.01	0.91	0.618	3.439	0.28	0.59	0.0
0.80	2.30	1.02	0.608	3.412	0.31	0.67	0.0
1.00	2.64	1.17	0.603	3.361	0.35	0.77	0.0
1.20	2.97	1.29	0.598	3.349	0.38	0.87	0.0
1.50	3.42	1.48	0.611	3.367	0.43	1.00	0.0
2.00	4.09	1.75	0.626	3.370	0.50	1.20	0.0
2.50	4.71	2.00	0.636	3.421	0.58	1.38	0.0
3.20	5.50	2.30	0.638	3.407	0.66	1.61	0.0
4.00	6.34	2.63	0.647	3.425	0.75	1.85	0.0
5.00	7.30	2.99	0.636	3.399	0.86	2.15	0.0
6.50	8.70	3.52	0.633	3.414	1.03	2.53	0.0
8.00	10.00	4.01	0.615	3.362	1.18	2.92	0.0
10.00	11.61	4.61	0.608	3.329	1.37	3.38	0.0
12.00	13.12	5.16	0.602	3.332	1.53	3.82	0.0
15.00	15.25	5.96	0.583	3.298	1.79	4.39	0.0
20.00	18.64	7.22	0.566	3.276	2.16	5.31	0.0
25.00	21.91	8.42	0.542	3.213	2.54	6.15	0.0
32.00	26.22	10.00	0.532	3.200	3.03	7.29	0.0
40.00	31.11	11.76	0.499	3.137	3.60	8.57	0.0
50.00	37.00	13.84	0.472	3.088	4.21	10.06	0.0
65.00	45.64	16.88	0.450	3.060	5.25	12.20	0.0
80.00	54.35	19.87	0.408	2.992	6.20	14.32	0.0
100.00	65.92	23.66	0.375	2.966	7.51	17.03	0.0
120.00	77.24	27.40	0.336	2.886	8.82	19.68	0.0
150.00	94.55	32.76	0.294	2.851	10.84	23.70	0.0
200.00	123.37	41.70	0.227	2.789	14.07	30.02	0.0
250.00	152.49	49.93	0.175	2.785	17.43	36.28	0.0
320.00	194.04	61.36	0.108	2.747	22.13	44.80	0.0
400.00	242.15	73.67	0.033	2.720	27.62	54.17	0.0
500.00	302.47	88.20	−0.046	2.742	34.49	65.50	0.0
650.00	393.11	108.10	−0.137	2.791	44.88	82.01	0.0
800.00	484.57	126.67	−0.222	2.841	55.34	97.66	0.0
1000.00	605.60	148.37	−0.329	2.952	68.91	117.23	0.0
1200.00	725.05	167.99	−0.406	3.053	82.88	135.80	0.0
1500.00	901.15	194.38	−0.512	3.240	102.63	160.59	0.0
2000.00	1184.74	229.03	−0.641	3.521	135.34	197.57	0.0
2500.00	1454.66	258.24	−0.765	3.853	165.47	229.20	0.0
3200.00	1813.43	289.53	−0.899	4.289	206.87	267.87	0.0
4000.00	2195.71	319.39	−1.010	4.676	250.95	304.41	0.0
5000.00	2643.71	347.70	−1.135	5.213	300.18	342.36	0.0
6500.00	3259.33	377.80	−1.279	5.902	372.32	389.37	0.0
8000.00	3826.47	401.70	−1.389	6.476	434.70	427.01	0.0

statistics analytically and in this mode is extremely fast. It is regularly updated and down-loadable from the web [17].

UT-Marlowe [21] differs from SRIM mainly in that it accepts 2D, single-crystalline targets and runs on a variety of platforms.

TABLE 7.6 Implant Statistics of In Implanted into Amorphous Si

E (keV)	R_p (nm)	σ (nm)	γ	β	$R_{p,lat}$ (nm)	σ_{lat} (nm)	Dose-Loss (%)
0.10	0.66	0.29	0.643	3.523	0.13	0.15	0.0
0.12	0.76	0.32	0.630	3.517	0.14	0.17	0.0
0.15	0.88	0.36	0.602	3.455	0.15	0.20	0.0
0.20	1.07	0.43	0.593	3.456	0.17	0.24	0.0
0.25	1.24	0.48	0.586	3.434	0.19	0.28	0.0
0.32	1.45	0.54	0.579	3.442	0.22	0.33	0.0
0.40	1.66	0.61	0.576	3.390	0.24	0.38	0.0
0.50	1.90	0.68	0.573	3.424	0.27	0.44	0.0
0.65	2.21	0.78	0.582	3.410	0.30	0.52	0.0
0.80	2.50	0.86	0.562	3.366	0.33	0.59	0.0
1.00	2.84	0.96	0.563	3.349	0.37	0.67	0.0
1.20	3.16	1.06	0.565	3.360	0.40	0.75	0.0
1.50	3.59	1.18	0.548	3.314	0.45	0.86	0.0
2.00	4.24	1.38	0.559	3.337	0.52	1.02	0.0
2.50	4.82	1.55	0.553	3.368	0.59	1.17	0.0
3.20	5.57	1.77	0.543	3.314	0.67	1.35	0.0
4.00	6.36	2.00	0.547	3.313	0.76	1.55	0.0
5.00	7.28	2.26	0.534	3.281	0.86	1.78	0.0
6.50	8.56	2.64	0.516	3.252	1.01	2.09	0.0
8.00	9.74	2.98	0.518	3.248	1.15	2.39	0.0
10.00	11.24	3.40	0.499	3.209	1.33	2.76	0.0
12.00	12.54	3.80	0.497	3.217	1.46	3.07	0.0
15.00	14.38	4.38	0.500	3.214	1.68	3.49	0.0
20.00	17.24	5.28	0.499	3.204	2.00	4.14	0.0
25.00	19.90	6.13	0.490	3.174	2.29	4.72	0.0
32.00	23.45	7.25	0.478	3.174	2.72	5.54	0.0
40.00	27.30	8.46	0.468	3.116	3.15	6.39	0.0
50.00	31.93	9.91	0.473	3.169	3.67	7.38	0.0
65.00	38.58	12.02	0.453	3.121	4.42	8.81	0.0
80.00	45.11	14.03	0.447	3.104	5.15	10.14	0.0
100.00	53.46	16.59	0.427	3.069	6.08	11.86	0.0
120.00	61.71	19.09	0.401	3.025	7.05	13.55	0.0
150.00	74.05	22.78	0.383	2.998	8.43	16.03	0.0
200.00	94.18	28.65	0.337	2.949	10.77	19.97	0.0
250.00	114.14	34.35	0.303	2.915	13.00	23.80	0.0
320.00	142.20	42.18	0.263	2.862	16.32	29.14	0.0
400.00	174.76	50.82	0.222	2.846	19.83	34.93	0.0
500.00	215.27	61.12	0.170	2.785	24.40	42.13	0.0
650.00	277.06	76.27	0.104	2.762	31.62	52.67	0.0
800.00	339.21	90.38	0.058	2.756	38.61	62.85	0.0
1000.00	423.03	108.82	−0.017	2.761	48.10	76.18	0.0
1200.00	506.43	126.17	−0.071	2.769	58.19	88.99	0.0
1500.00	632.92	149.87	−0.144	2.808	71.69	107.31	0.0
2000.00	842.53	185.55	−0.255	2.895	95.98	136.58	0.0
2500.00	1050.50	216.38	−0.335	2.974	119.34	163.35	0.0
3200.00	1333.66	254.77	−0.431	3.120	151.69	197.18	0.0
4000.00	1649.53	291.04	−0.537	3.312	187.11	231.55	0.0
5000.00	2029.09	328.73	−0.630	3.516	231.33	270.74	0.0
6500.00	2568.46	373.24	−0.747	3.808	293.19	320.86	0.0
8000.00	3076.54	409.96	−0.849	4.115	349.16	362.71	0.0

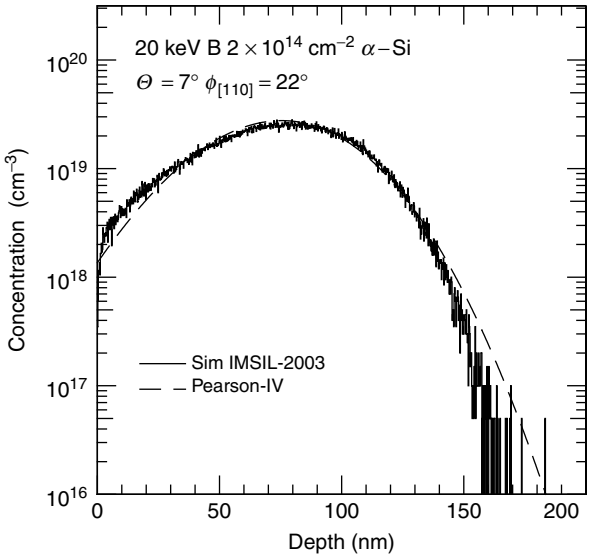


FIGURE 7.11 Concentration of B as a function of depth for a 20 keV B 2×10^{14} cm^{-2} implant into amorphous Si, simulated by a Monte Carlo (MC) technique (solid line), and the B concentration obtained from the Pearson-IV distribution using the moments extracted from the simulated depth profile (dashed line).

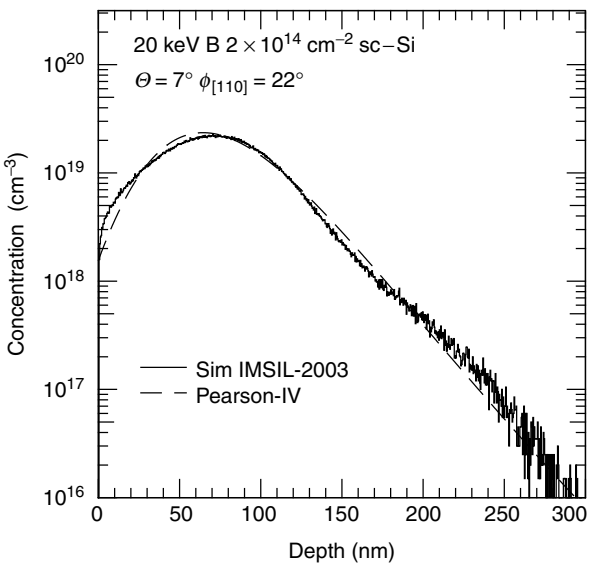


FIGURE 7.12 Concentration of B as a function of depth for a 20 keV B 2×10^{14} cm^{-2} implant into single-crystalline Si with a tilt angle of 7° and an azimuthal angle of 22°, simulated by IMSIL (solid line), and the B concentration obtained from the Pearson-IV distribution (dashed line) using the moments extracted from the simulated depth profile.

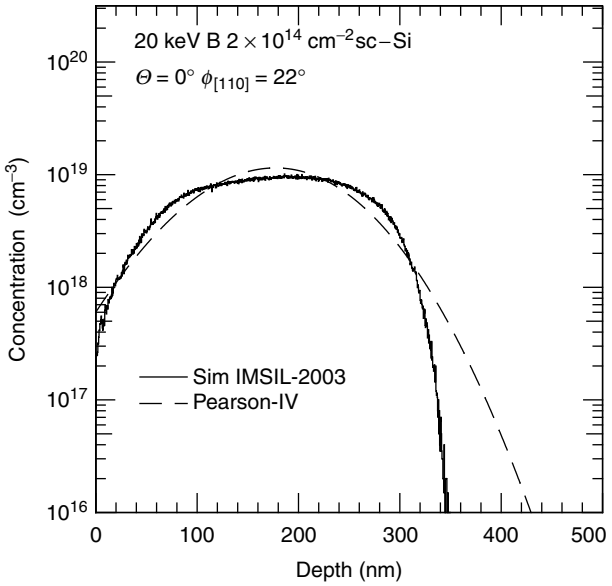


FIGURE 7.13 Concentration of B as a function of depth for a 20 keV B 2×10^{14} cm^{-2} implant into single-crystalline Si perfectly aligned with a ⟨100⟩ axis, i.e., with a tilt angle of 0° and an azimuthal angle of 0°, simulated by a IMSIL (solid line), and the B concentration obtained from the Pearson-IV distribution (dashed line) using the moments extracted from the simulated depth profile.

IMSIL [20] in addition accepts non-planar target geometries and multi-layers. It is available for simulation of most of the examples discussed here.

7.2.4 Channeling

In an amorphous target the atoms are assumed to be randomly distributed in a way that conforms to the overall packing density, local atomic coordination, and atom size constraints, in contrast to a single crystal, where the atoms are packed in a long-range ordered arrangement. For Si, Ge, and III–V's this is the diamond cubic lattice structure. The result of this symmetry is that an ion in a beam aligned close to a major crystallographic axis "sees" the target as a collection of atoms aligned up along "rows" with open space ("channels") in between the rows. As long as the angle between the incident beam and the major crystallographic axis does not exceed a critical value the incident ions will be "steered" by the atoms in the rows through a series of correlated small-angle collisions. The ions are said to be "channeled" and can travel long distances without encountering a target atom and undergoing a nuclear collision (and a nuclear energy loss). In the diamond-cubic lattice the ⟨100⟩, ⟨111⟩, and ⟨110⟩ directions provide such paths. In addition there are highly symmetric planar directions, such as the {100}, {111}, and {110} planes which also provide ion channeling opportunities.

An appropriately oriented ion beam will either be directly aligned along such channel directions or allow a high fraction of incident ions to be scattered into a channel. Channeled ions will penetrate deeply into the silicon since overall stopping is determined primarily by the electronic stopping component. Channeling of energetic ions is well understood and described [22–25].

The propensity for ions to be scattered into and remain captured in a channel is a function of incident angle, ion mass, and energy. Quantitatively it is described by the "critical angle," ψ_c. Qualitatively, if the angle between the beam and the major crystallographic axis exceeds ψ_c ions will, in general, no longer be steered into the channels and the target appears amorphous (Table 7.7).

TABLE 7.7 Implant Statistics of Sb Implanted into Amorphous Si

E (keV)	R_p (nm)	σ (nm)	γ	β	$R_{p,lat}$ (nm)	σ_{lat} (nm)	Dose-Loss (%)
0.10	0.67	0.29	0.607	3.466	0.13	0.14	0.0
0.12	0.76	0.32	0.582	3.451	0.14	0.17	0.0
0.15	0.89	0.36	0.557	3.387	0.15	0.19	0.0
0.20	1.08	0.42	0.547	3.402	0.17	0.24	0.0
0.25	1.25	0.48	0.538	3.388	0.19	0.28	0.0
0.32	1.45	0.54	0.531	3.362	0.22	0.32	0.0
0.40	1.66	0.60	0.536	3.373	0.24	0.37	0.0
0.50	1.90	0.67	0.536	3.402	0.26	0.43	0.0
0.65	2.21	0.77	0.533	3.355	0.30	0.50	0.0
0.80	2.50	0.85	0.531	3.340	0.33	0.57	0.0
1.00	2.84	0.95	0.531	3.314	0.37	0.65	0.0
1.20	3.15	1.04	0.545	3.370	0.40	0.73	0.0
1.50	3.58	1.16	0.531	3.311	0.45	0.83	0.0
2.00	4.22	1.35	0.509	3.252	0.52	0.99	0.0
2.50	4.80	1.52	0.531	3.305	0.59	1.13	0.0
3.20	5.54	1.73	0.514	3.245	0.67	1.31	0.0
4.00	6.32	1.95	0.509	3.250	0.75	1.50	0.0
5.00	7.21	2.20	0.506	3.229	0.86	1.71	0.0
6.50	8.47	2.56	0.513	3.272	1.00	2.01	0.0
8.00	9.64	2.89	0.493	3.203	1.14	2.30	0.0
10.00	11.10	3.29	0.504	3.233	1.30	2.65	0.0
12.00	12.36	3.66	0.493	3.204	1.44	2.94	0.0
15.00	14.17	4.22	0.490	3.198	1.67	3.34	0.0
20.00	16.89	5.07	0.485	3.171	1.97	3.93	0.0
25.00	19.48	5.86	0.477	3.170	2.26	4.51	0.0
32.00	22.89	6.91	0.469	3.163	2.66	5.25	0.0
40.00	26.56	8.08	0.470	3.154	3.08	6.03	0.0
50.00	31.05	9.42	0.459	3.149	3.56	6.97	0.0
65.00	37.34	11.34	0.439	3.090	4.30	8.24	0.0
80.00	43.51	13.21	0.428	3.087	4.98	9.50	0.0
100.00	51.35	15.60	0.408	3.047	5.89	11.10	0.0
120.00	59.14	17.79	0.387	3.027	6.77	12.63	0.0
150.00	70.55	21.21	0.377	2.988	8.07	14.84	0.0
200.00	89.28	26.42	0.323	2.926	10.21	18.43	0.0
250.00	107.78	31.50	0.298	2.918	12.40	21.85	0.0
320.00	133.56	38.54	0.255	2.881	15.32	26.40	0.0
400.00	162.88	45.90	0.208	2.828	18.62	31.61	0.0
500.00	199.71	55.11	0.166	2.806	22.84	37.81	0.0
650.00	255.02	67.90	0.102	2.783	28.99	46.78	0.0
800.00	310.14	80.25	0.035	2.751	35.19	55.61	0.0
1000.00	383.38	95.50	−0.021	2.766	43.50	66.70	0.0
1200.00	457.35	110.00	−0.079	2.790	51.94	77.34	0.0
1500.00	567.02	129.26	−0.145	2.811	64.53	92.62	0.0
2000.00	747.44	158.73	−0.264	2.902	85.06	115.97	0.0
2500.00	924.14	183.27	−0.333	2.979	104.95	138.20	0.0
3200.00	1164.09	213.21	−0.436	3.137	132.68	165.02	0.0
4000.00	1428.39	241.84	−0.523	3.302	161.76	192.83	0.0
5000.00	1744.44	272.15	−0.632	3.544	199.05	222.42	0.0
6500.00	2190.23	305.58	−0.743	3.833	250.62	260.69	0.0
8000.00	2608.65	332.07	−0.814	4.072	297.27	293.91	0.0

The critical angle for axial channeling can be written as [22,23,26]

$$\psi_e = F\sqrt{\frac{Z_1 Z_2 e^2}{4\pi\varepsilon_0 E d}} \tag{7.30}$$

TABLE 7.8 Values of the Potential Function, F, Defined in Equation 7.31, for the Implantation of Common Dopants into Si at Room-Temperature

Specie	F
B	1.483
P	1.385
As	1.294
In	1.241
Sb	1.236

The potential function F is given by

$$F = \sqrt{\ln\left[\frac{3a^2}{u_1^2} + 1\right]}, \tag{7.31}$$

where a is the screening distance suggested by Firsov [27]

$$a = \frac{0.8853 a_0}{\left(Z_1^{1/2} + Z_2^{1/2}\right)^{2/3}}, \tag{7.32}$$

and d is the distance between atoms along the atomic rows. The parameter u_1 is the root-mean-square vibrational amplitude of the target atoms; for silicon at room-temperature $u_1 = 8.7$ pm [28]. The potential function F is tabulated in Table 7.8 for some common dopants. Table 7.9 lists values of d for some selected major crystallographic axes in the diamond cubic lattice. The critical angles are shown as a function of energy in Figure 7.14.

We note that Equation 7.30 should be considered as a guide only. Accurate determination of ψ_c requires a MC simulation. Nevertheless, it is quite obvious that ψ_c can reach some rather large values for low energy ion beams.

It is now clear why the implant at a tilt of 7° in Figure 7.12 failed to give a profile like that into an amorphous target: The critical angle for 20 keV B is of the order 8° and the chosen tilt angle is simply not sufficient to make the target appear amorphous.

Achieving an implant profile without appreciable channeling is of practical importance to avoid that slight differences in beam orientation across the wafer result in radically different implant profiles. Ziegler and Lever [29] have mapped the channels in Si by measuring the backscattering yield of He and N. Lever and Brannon [30,31] have produced a similar map by simulating the mean free path of 100 keV B atoms in Si as a function of incident angles. Figure 7.15 is a different view of the same problem; there we have plotted the junction depth, defined as the depth at which the dopant concentration reaches 1×10^{18} cm^{-3}, for a 100 keV B implant of dose 5×10^{14} cm^{-2}. For this IMSIL simulation damage does not accumulate.

Based on those kinds of simulations and experiments, "magic" angles of $\Theta = 7°$ and $\Phi = 22°$ are often used. However, is apparent from the increase in critical angle depicted in Figure 7.14, for low energies this

TABLE 7.9 Distance along Rows, d, in Units of the Lattice Constant, d_0, for Some Selected Major Crystallographic Axes in the Diamond Cubic Lattice. For Si $d_0 = 0.5431$ nm

Axis	d/d_0
⟨100⟩	1
⟨110⟩	$1/\sqrt{2}$
⟨111⟩	$\sqrt{3/4}$, $\sqrt[3]{3/4}$

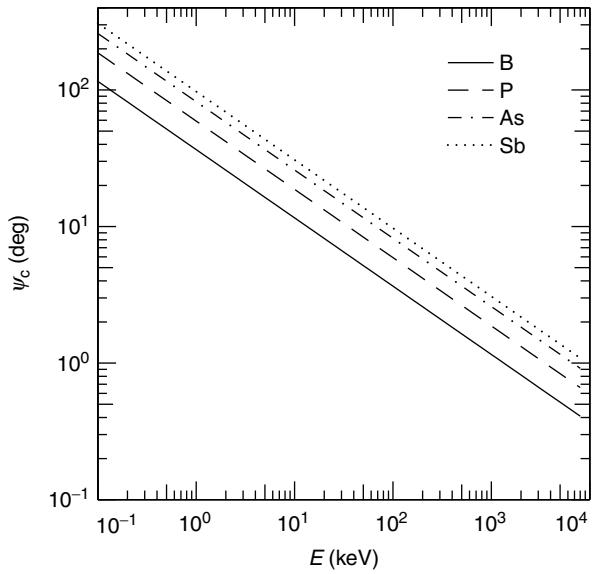


FIGURE 7.14 Critical angle as a function of energy for the implantation of common dopants into Si at room-temperature.

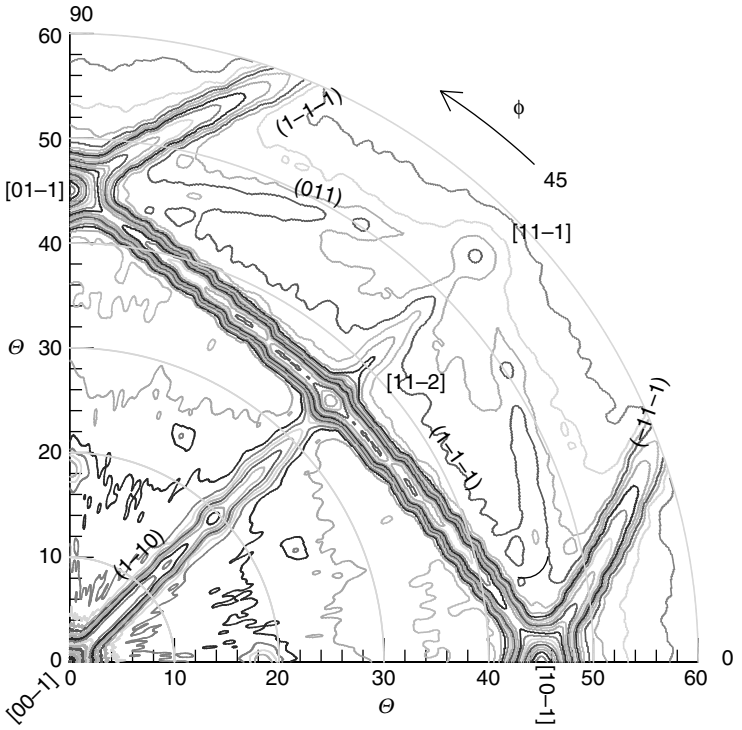


FIGURE 7.15 Junction depth, defined as the depth at which the dopant concentration reaches 1×10^{18} cm^{-3}, for a 100 keV B implant of dose 5×10^{14} cm^{-2}. Major crystallographic axes or planes are characterized by closely spaced contour lines (IMSIL simulation without damage accumulation).

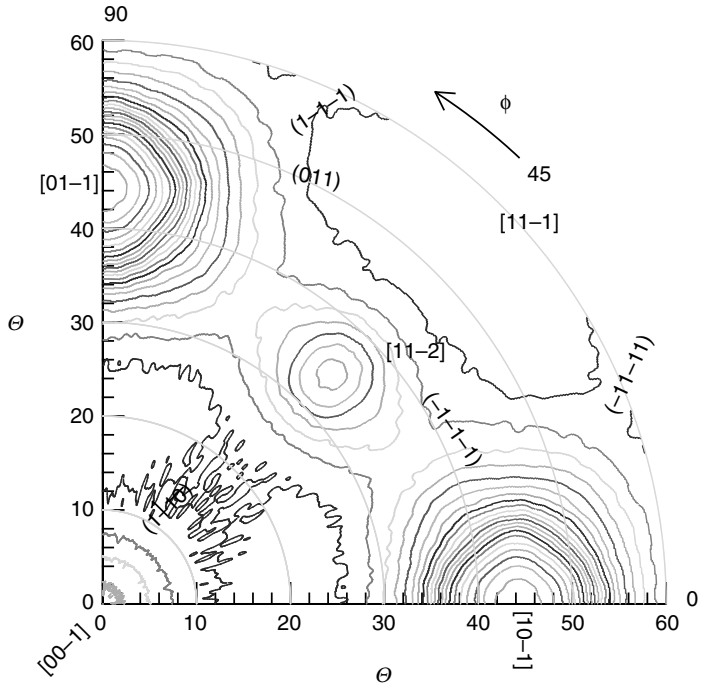


FIGURE 7.16 Junction depth, defined as the depth at which the dopant concentration reaches 1×10^{18} cm^{-3}, for a 0.5 keV B implant of dose 5×10^{13} cm^{-2}. Major crystallographic axes or planes are characterized by closely spaced contour lines (IMSIL simulation without damage accumulation).

becomes questionable. Indeed it has been reported [32] that the channeling tail of a 0.5 keV B implant is almost independent of tilt angle.

Figure 7.16 shows the same kind of map as Figure 7.15, but now for 0.5 keV B at a dose of 5×10^{13} cm^{-2}. As one could have guessed from Figure 7.14 the major axes have large acceptance angles for channeling. Interestingly, planar channeling is absent and so it is channeling into high-order index axes [31].

While it may seem that at those low energies essentially all incident angles lead to more or less channeling this is only true if there is no screen oxide present. However, in practice, even in the absence of a deliberate screen oxide, there will always be a native oxide present. For a 100 keV B implant the influence of the native oxide can safely be ignored. This is not true for a 0.5 keV B implant. The thickness of the native oxide is of order 1 nm, a large fraction of the range of a 0.5 keV B. Hence the incoming atoms will undergo significant scattering in the amorphous oxide and the beam entering the crystal will have a large divergence.

This is illustrated in Figure 7.17, where again the junction depth is plotted, as defined previously for Figure 7.15 and Figure 7.16, for a 0.5 keV B implant of dose 5×10^{13} cm^{-2}, but now in the presence of 1.4 nm native oxide. Channeling is dramatically reduced to the point where it has become hard to make out any structure in the contour plot.

7.2.5 Defects

Ion implantation is an inherently violent process. The energy that the incoming ion sheds, while it slows down is transferred to the electrons and nuclei of the target. The latter may become dislodged, forming a Frenkel-pair, or, if sufficient energy is transferred, become a "knock-on" and displace further target atoms

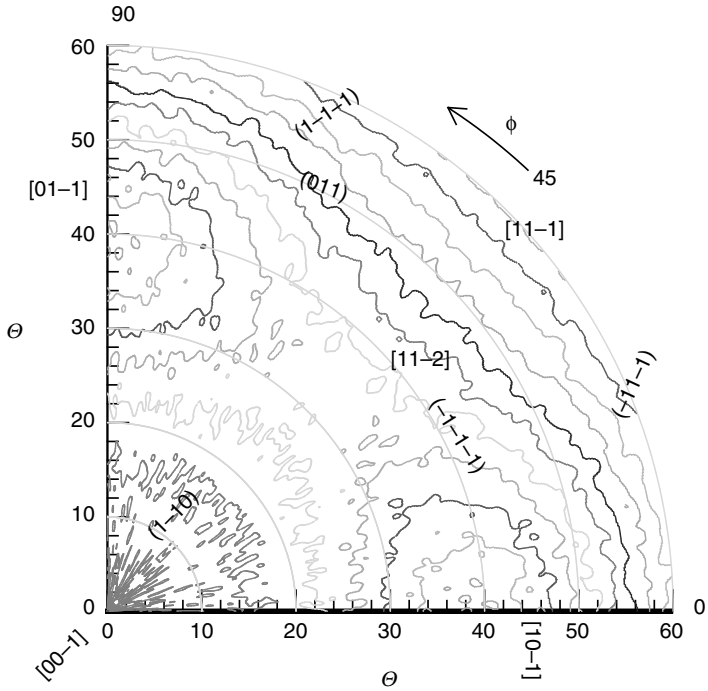


FIGURE 7.17 Junction depth, defined as the depth at which the dopant concentration reaches 1×10^{18} cm^{-3}, for a 0.5 keV B implant of dose 5×10^{13} cm^{-2}, in the presence of 1.4 nm of native oxide (IMSIL simulation without damage accumulation).

("collision cascade"). The ensuing damage is usually described in the form of Frenkel-pairs, interstitials and vacancies of the target atoms. They are typically mobile and interact among themselves and with each other. Interstitials may form small clusters, or larger rod-like defects or dislocation loops [33]; vacancies agglomerate to clusters or voids [34]; the interaction of interstitials with vacancies is usually assumed to lead to recombination [35].

Understanding and modeling of as-implanted damage has remained a difficult problem [36] and continues to be listed as one of the difficult challenges of Modeling and Simulation in the International Technology Roadmap for Semiconductors (ITRS) [37]. This is due to the many parameters that govern defect formation, defect annealing, clustering and amorphization.

7.2.5.1 The Standard Model of Defect Production

The standard treatment of radiation-induced damage follows the seminal work of Kinchin and Pease [38] who introduced the concept of a "displacement energy," E_d. A stationary target atom is displaced from its lattice site, and may itself displace other, stationary target atoms, if the energy transferred during the primary collision exceeds E_d. In Si $E_d = 15$ eV is usually assumed. Using statistical arguments and an unscreened potential Kinchin and Pease show that the number of target atoms displaced by an incident ion is given by [38]

$$N_d = \frac{1}{2}\left\{1 + \ln\left[\frac{4M_1M_2}{(M_1+M_2)^2}\,\frac{E}{2E_d}\right]\right\} \tag{7.33}$$

The concept of the displacement energy is straightforward to implement into a MC program and the depth distribution of displaced atoms may be calculated by following each recoil with energy above E_d.

Assuming that the target turns amorphous if a certain density of displaced atoms [39–41] or a critical energy density [42–45] is exceeded allows a relatively good prediction of amorphous layer thickness.

The standard model can be improved by explicitly measuring the depth distribution of damage, e.g., by ion channeling, and fitting empirical parameters. Hobler [46] has reported on the use of an empirical factor f_{rec} that depends only on the ion species. Defect recombination (dynamic annealing) within the recoil cascade may be described by a value of $f_{rec} < 1$; overlapping or dense cascades with their more efficient defect generation may lead to $f_{rec} > 1$. Amorphization of a region in the target is assumed to take place if the point defect density exceeds a critical value.

From Figure 7.4, we see that the nuclear stopping power depends on energy and exhibits maximum at certain energy. Further, at sufficiently high energy, electronic stopping (Figure 7.2) exceeds nuclear stopping. Therefore, when an ion initially impinges on the target at high energy, stopping is predominantly composed of electronic and non-displacement nuclear stopping. However, as the ion is slowed through these non-damaging processes, an energy is reached where target displacement collisions become predominant and significant target damage is produced. After further slowing, the ion has lost sufficient energy that only collisions with small impact parameter can produce displacements and damage production drops off again. We therefore expect a damage peak at a shallower depth than the maximum of the dopant distribution. Figure 7.18 illustrates this by a simulation of 5 keV P implant. The dose of 1×10^{13} cm^{-2} has been chosen to avoid amorphization. The vertical solid line indicates the maximum in the P concentration, clearly deeper than the maximum in the defect concentration.

Although interstitial and vacancy concentration profiles appear identically a more careful analysis shows that this is not the case. The interstitial profile is slightly shifted to larger depths due to the creation of a static vacancy and a mobile knock-on that moves into the target. We define the *excess* interstitial concentration, N_I^{excess}, as

$$N_I^{excess}(x) = N_I(x) - N_V(x), \tag{7.34a}$$

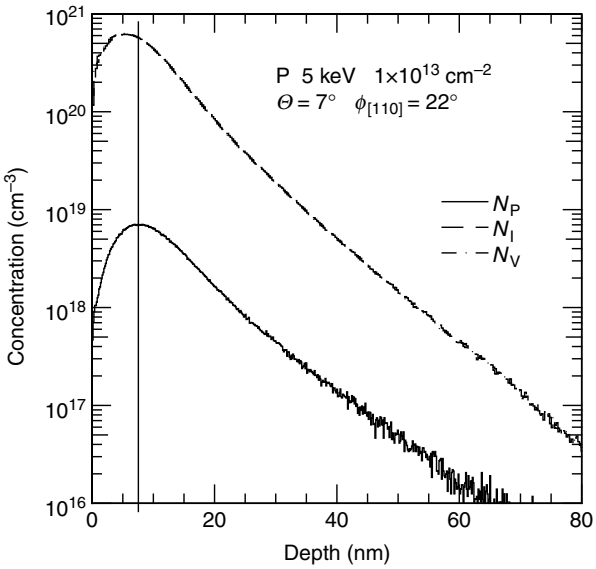


FIGURE 7.18 Damage concentration (interstitial, N_I, and vacancy, N_V) and P concentration, N_P, a function of depth for a 5 keV 1×10^{13} cm^{-2} P implant (IMSIL simulation). The vertical solid line indicates the depth of the maximum P concentration.

and the excess vacancy concentration, N_V^{excess}, as

$$N_V^{excess}(x) = N_V(x) - N_I(x), \tag{7.34b}$$

where N_I and N_V are the volume concentrations of Si self-interstitials and vacancies, respectively, and x is the depth into the wafer of a particular bin. Volume concentrations in a MC calculation are arrived at by dividing the number of atoms or defects by a pre-defined elementary volume, a bin. We will call the linear dimension of this volume the bin-width. Figure 7.19 shows the excess interstitial concentration (dashed line) and the excess vacancy concentration (dotted line) for a 5 keV 1×10^{13} cm^{-2} P implant The bin-width is 0.14 nm; hence this is equivalent to assuming that vacancies and interstitials recombine if they are closer than 0.14 nm.

As Figure 7.19 shows the maximum in the dopant concentration (solid line) separates a vacancy-rich region close to the surface from an interstitial-rich region in the bulk. The evolution upon annealing of those defect distributions by diffusion, agglomeration, and recombination is of great importance for the final dopant profile and further discussed in Section 7.3.5.

The separations between damage and dopant peaks and the extent of the vacancy-rich region can be increased by increasing the projectile energy [47,48]. Figure 7.20 shows the excess interstitial (dashed line) and excess vacancy (dotted line) concentration, together with the dopant concentration (solid line) as a function of depth for a 500 keV 1×10^{13} cm^{-2} P implant (IMSIL simulation). Here the bin-width and hence the implicit recombination distance was 13 nm.

Implanting at high-energy into a silicon-on-insulator (SOI) structure with an appropriate Si film thickness allows placement of the bulk of the interstitials into the buried oxide film where they will not be able to interact with the vacancies. In this fashion samples containing predominantly vacancies can be produced [49] ("Vacancy Implanter") that allow to study the behavior of vacancies during annealing.

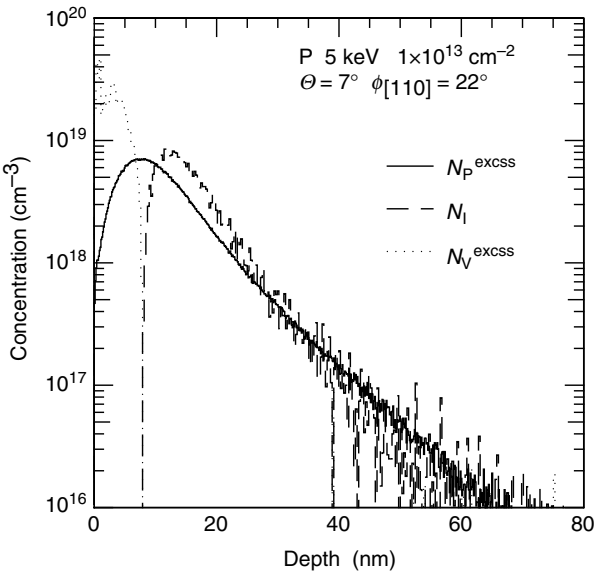


FIGURE 7.19 Excess interstitial concentration, N_I^{excess}, (dashed line) and excess vacancy concentration, N_V^{excess}, (dotted line) together with the P concentration, N_P, (solid line) as a function of depth for a 5 keV 1×10^{13} cm^{-2} P implant (IMSIL simulation).

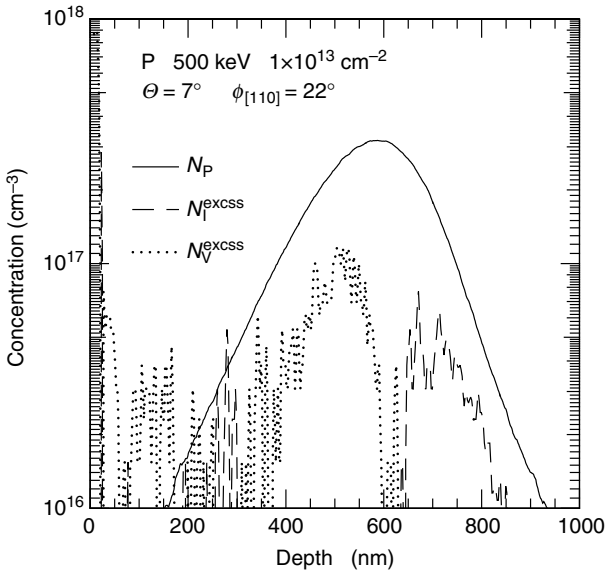


FIGURE 7.20 Excess interstitial concentration, N_I^{excess}, (dashed line) and excess vacancy concentration, N_V^{excess} (dotted line), together with the P concentration, N_P, (solid line) as a function of depth for a 500 keV 1×10^{13} cm^{-2} P implant (IMSIL simulation).

7.2.5.2 Amorphization

Damage accumulation and eventual amorphization results from the interplay between displacements of target atoms and dynamic annealing. The critical dose for amorphization, Φ_α, therefore depends not only on projectile species and energy but also on target temperature during implant and dose-rate. When defect production and dynamic annealing are approximately balanced, damage accumulation becomes very non-linear with dose [50]. At low doses the damage level is relatively low; around a "critical dose" damage increases rapidly; at high doses damage saturates. A similar behavior is observed as a function of implant temperature at fixed dose and dose-rate, [51] as illustrated in Figure 7.21 for 1 MeV Si implant of dose 1×10^{15} cm^{-2}.

The data can be well described [51] by

$$Y_{\text{norm}} = c_1 + \frac{c_2}{1 + \exp\left[c_3(T - T_c)\right]}, \tag{7.35}$$

where Y_{norm} is the normalized damage and c_i, $i=1, 2, 3$ are constants. The critical temperature T_c is related to the dose rate by an Arrhenius relationship; for 1 MeV. 1×10^{15} cm^{-2} Si implant in Figure 7.21 the activation energy is 0.9 eV [51].

For implants well below or well above T_c, i.e., in the flat parts of the curves in Figure 7.21, amorphization is independent of temperature and hence dose-rate. However, in the steep part of the curve the amount of damage produced depends very much on implant temperature and rate. If the operating point of an implanter is chosen in this regime implant temperature and dose-rate need to be carefully controlled.

While only weakly dependent on dose-rate, T_c has a strong dependence on ion mass. Goldberg et al. have measured [50] T_c for the species listed in Table 7.10 and for dose-rates ranging from 5×10^{11} cm^{-2} s^{-1} to 5×10^{13} cm^{-2} s^{-1}. Table 7.10 gives T_c at a dose-rate of 3.2×10^{12} cm^{-2} s^{-1}. Note that the lighter elements C and Si both have critical temperatures of the same order as typical implant temperatures.

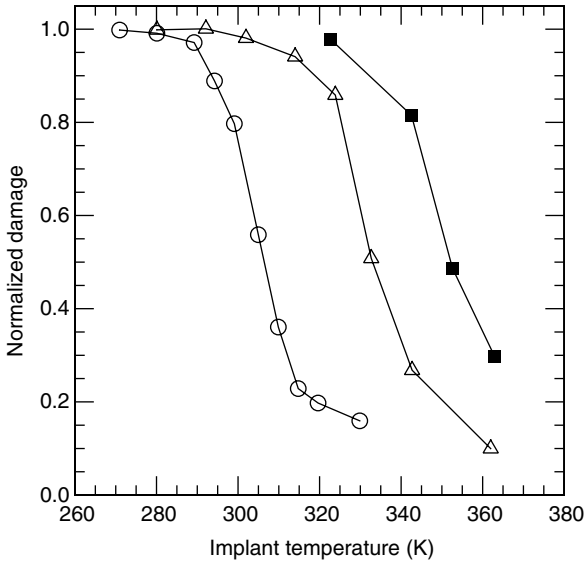


FIGURE 7.21 Normalized damage as a function of implant temperature for a 1 MeV Si implant of dose 1×10^{15} cm^{-2} at dose-rates of 2.88×10^{11} cm^{-2} s^{-1} (circles), 3.57×10^{12} cm^{-2} s^{-1} (triangles), and 2.63×10^{13} cm^{-2} s^{-1} (squares), respectively. (Data from Schultz, P. J., et al., *Phys. Rev. B*, 44, 9118, 1991.)

In the standard model of defect production the target is treated as amorphous if the point defect concentration exceeds a certain critical value, N_{CPDC}. As Figure 7.22 illustrates for a simulated 5 keV Si implant of dose 5×10^{14} cm^{-2}, the resulting position of the amorphous–crystalline (a–c) interface is not very sensitive to the exact value of N_{CPDC}. In the example of Figure 7.22, 100% change of N_{CPDC} from 1.5×10^{22} cm^{-3} (horizontal, dotted line) to 3×10^{22} cm^{-3} (horizontal, dash-dotted line) moves the a–c interface position from 16 to 12 nm, change of -25% only. This is because the shape of the defect profile is approximately Gaussian and rather steep in the vicinity of the a–c interface. The standard model therefore gives frequently quite reasonable values for the thickness of the amorphous layer.

For Si values for N_{CPDC} in the literature vary between 5 and 50% of the atomic density of Si [36]. Hobler and Otto have extracted [36] N_{CPDC} from published, measured amorphous layer thicknesses, and simulations using the binary-collision approximation. Values vary by one order of magnitude with no systematic dependence on ion species or energy. The authors report a weak correlation on dose and the observation of a dependence on dose-rate; nevertheless, the large scatter in the data prevented extraction of a reliable model.

TABLE 7.10 Critical Temperature of Amorphization, T_c, at a Dose-Rate of 3.2×10^{12} cm^{-2} s^{-1} for a Variety of Species. Also Shown Is the Activation Energy

Species	T_c (K)	E_a (eV)
C	290	0.70
Si	341	0.79
Ar	401	0.99
Ge	430	1.11
Kr	500	1.26
Xe	582	1.69

Source: From Goldberg, R.D., Williams, J.S., and Elliman, R.G., *Nucl. Instrum. Methods Phys. Res. B*, 106, 242, 1995.

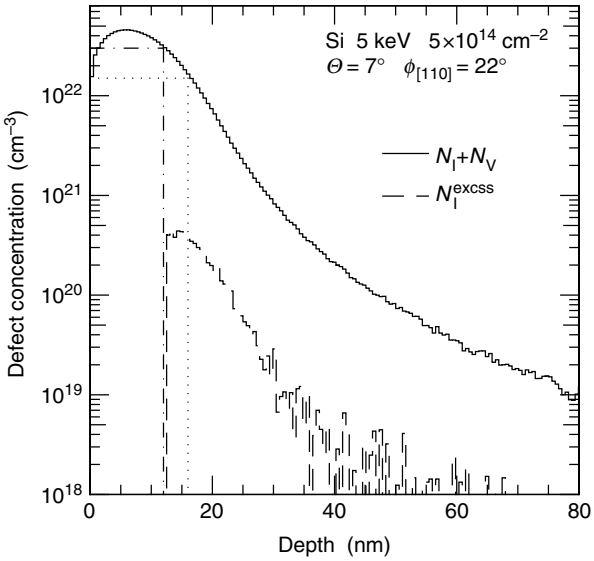


FIGURE 7.22 Total defect concentration (interstitials and vacancies) (solid line) and excess (net) interstitial concentration (dashed line) as a function of depth for a 5 keV Si implant of dose 5×10^{14} cm^{-2} (IMSIL simulation). The horizontal dash-dotted and dotted lines indicate a critical point defect concentration of $N_{CPDC}=3\times10^{22}$ cm^{-3} and of $N_{CPDC}=1.\ 5\times10^{22}$ cm^{-3}, respectively, leading to an amorphous–crystalline interface as indicated by the vertical dash-dotted and dotted lines, respectively.

After a thermal treatment the amorphous layer recrystallizes and the usual assumption is that the damage in it is completely removed. The damage beyond the a–c interface recombines until all vacancies are consumed. As was discussed above, this region of the target has an interstitial excess and after recombination a finite amount of interstitials remain (dashed line in Figure 7.22), most of which eventually agglomerate, forming end-of-range (EOR) defects. Contrary to the position of the a–c interface, the concentration of interstitials available for EOR defects depends rather sensitively on the value of N_{CPDC}. In Figure 7.22 the two values of N_{CPDC}, differing by 100%, give an integrated excess (net) interstitial concentration of 1.9×10^{14} cm^{-2} and 3.7×10^{14} cm^{-2}, respectively, i.e., a change of 95%. The residual damage has an uncertainty associated with it that is of the same order as the uncertainty in N_{CPDC} [38].

The excess interstitials beyond the a–c interface have a significant impact on the final dopant profile after annealing. In general, the standard defect model incorporated into binary-collision simulators does not satisfactorily predict this crucial value. Only simulations that take explicitly the energetics of defect formation, agglomeration, and dissolution into account can be expected to be predictive. Pelaz et al. have recently proposed such an approach implemented [52].

7.2.6 List of Symbols Used in This Section

a	screening length
a_{OR}	Oen–Robinson screening length
a_0	Bohr radius ($a_0=5.293\times10^{-11}$ m)
a_U	screening length in the ZBL-theory of nuclear stopping
$a_{U,i}$	individual components' screening length of the universal screening function of the ZBL-theory of nuclear stopping
b	impact parameter

b_{max}	maximum value of the impact parameter
c	weighting exponent in the generalized theory of electronic stopping
c_i	parameters used in the modeling of ion-beam-induced amorphization; $i=1, 2, 3$
d	distance between atoms along atomic rows
d_0	latice constant (in Si $d_0=0.5431$ nm)
$\Delta E_{e,nl}$	non-local energy loss due to electronic stopping
$\Delta E_{e,loc}$	local energy loss due to electronic stopping
E	ion or projectile energy
E_d	displacement energy (in Si $E_d=15$ eV)
e	elementary charge ($e=1.60122\times10^{-19}$ C)
F	potential function used in the calculation of the channeling critical angle
f_s	screening function
f_U	universal screening function of the ZBL-theory of nuclear stopping
$f_{U,i}$	individual components' weighting factor of the universal screening function of the ZBL-theory of nuclear stopping
$\hbar$	Planck constant ($\hbar=6.6262\times10^{-34}$ Nms)
I	average excitation potential
k	stopping power proportionality constant in the generalized theory of electronic stopping
k_L	stopping power proportionality constant in the LSS theory of electronic stopping
m, ν, a, λ	Pearson-IV distribution parameters
M_1	atomic mass of incident ion or projectile ion
M_2	atomic mass of target atom
m_e	electron mass ($m_e=9.1095\times10^{-31}$ kg)
N	atomic density (in Si $N=4.99$ cm^{-3})
N_A	volume concentration of dopant atom A
N_I	volume concentration of Si self-interstitials
N_I^{excess}	excess or net Si self-interstitial concentration
N_P	volume concentration of phosphorus
N_V	volume concentration of Si vacancies
N_V^{excess}	excess or net Si vacancy concentration
q	non-local fraction exponent
R	range
R_p	projected range
$R_{p,lat}$	lateral, projected range
r	stopping power exponent in the generalized theory of electronic stopping
S	stopping power
S_e	electronic stopping power
S_{eB}	electronic stopping power according to the Bethe–Bloch theory of electronic stopping
S_{eL}	electronic stopping power according to the LSS theory of electronic stopping
S_{er}	electronic stopping power according to the generalized theory of electronic stopping
S_n	nuclear stopping power
T	energy transfer in a collision; alternatively temperature
T_c	critical temperature used in the modeling of ion-beam-induced amorphization
T_{max}	maximum energy transfer in a collision
u_1	root-mean-square vibrational amplitude (in Si at room-temperature $u_1=8.7$ pm)
V	interatomic potential
v	velocity
v_0	Bohr velocity ($v_0=2.187\times10^6$ m/s)
v_1	velocity of incident ion or projectile ion before collision
v_1'	velocity of incident ion or projectile ion after collision
v_2	velocity of target atom before collision
v_2'	velocity of target atom after collision
v_c	critical velocity

x	depth
x_{loc}	local fraction of electronic stopping
x_{nl}	non-local fraction of electronic stopping
Y_{norm}	normalized damage
y_{nl}	non-local fraction proportionality constant
Z_1	atomic number of incident ion or projectile ion
Z_2	atomic number of target atom
β	kurtosis
γ	skewness
ε	stopping cross-section
ε_0	vacuum permittivity ($\varepsilon_0 = 8.8542 \times 10^{-12}$ F/m)
Θ	scattering angle; alternatively tilt angle of incident ion beam with respect to wafer normal
μ_i	i-th moment of a distribution, $i = 1, 2, \ldots, 4$
σ_{lat}	lateral, projected straggle
σ_p, σ	straggle, projected straggle
$d\sigma$	differential cross-section
Φ	implanted dose per unit area; alternatively azimuthal angle of incident ion beam in the wafer frame of reference
ψ_c	channeling critical angle
$d\Omega$	differential solid angle

7.3 Applications of Ion Implantation

7.3.1 Introduction

Ion implantation is firmly entrenched in production processes for leading edge CMOS integrated circuit fabrication. The ion implantation process is highly flexible in the selection of dopant species, in choosing the spatial location within the device, and in providing subtle concentration profile control. It can adapt to changes required by other process advances, enabling rapid introduction of new integrated process technology. It provides many degrees of freedom in processing, allowing precise optimization of performance tradeoffs as new circuit technologies with scaled device dimensions that are introduced.

Basic CMOS processes usually use fifteen to seventeen ion implants per wafer; current leading edge CMOS processes use twenty to twenty three implants, and some specialized CMOS circuits (e.g., flash memory) use up to thirty implantation steps. Figure 7.23 shows the doped regions of an advanced CMOS structure. Virtually all doping in modern CMOS devices is accomplished by ion implantation; no other technique offers comparable process control and repeatability for both the amount and position of the doping. Metal oxide semiconductor field effect transistor (MOSFET) doping requirements span several orders of magnitude in both dose and energy for multiple species, as shown in Figure 7.24. Modern ion implantation equipment has been refined such that two to three carefully designed system architectures are capable of covering all of these applications. A detailed description of all mainstream applications [53] is beyond the scope of this chapter; instead we describe the two leading edge applications at either end of the energy spectrum.

7.3.2 Buried Layers

Buried layers are highly doped (10^{19} cm^{-3}) regions implanted 2–4 μm beneath the silicon surface, usually with no photomask present ("Latchup/ESD protection in Figure 7.24"). Buried layers can greatly enhance device performance on bulk substrates, and have the potential for replacement of epi silicon substrates [54,55]. The doping and EOR damage in the buried layer region reduce the concentrations of metals and oxygen in the surface region in undenuded wafers down to those in denuded silicon, and almost down to the levels in epi layers. Additionally, the EOR damage getters defects induced by subsequent implants, reducing leakage currents [56–60]. Despite a higher cost, epi substrates are not

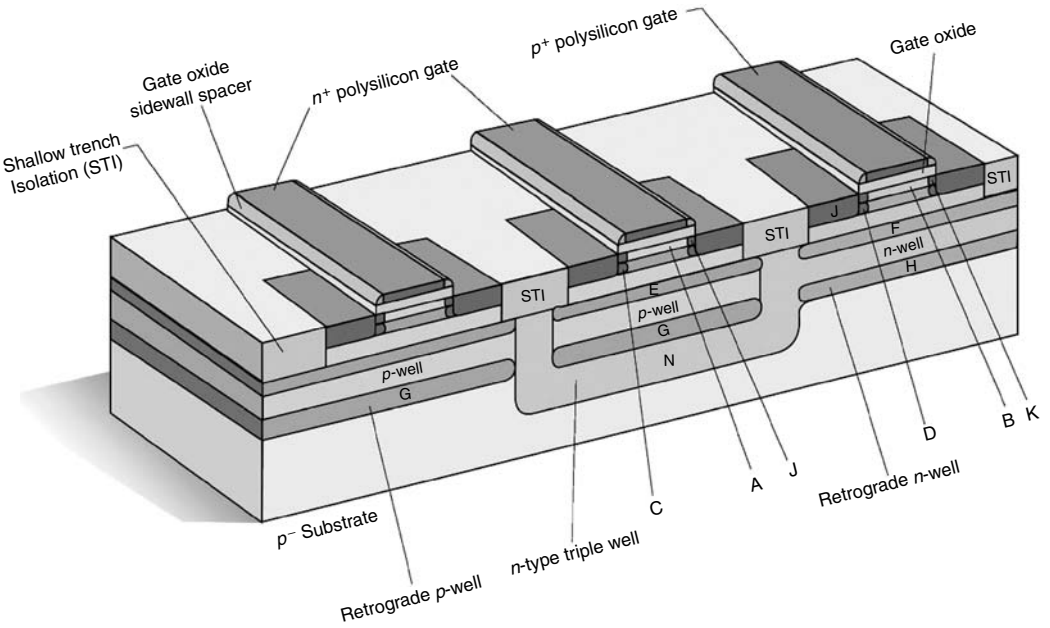


FIGURE 7.23 Implanted regions in advanced devices. A, *n*-channel metal oxide silicon (NMOS) channel threshold voltage adjust (*p*-type); B, PMOS channel threshold voltage adjust (*n*-type); C, NMOS punchthrough stop (*p*-type); D, PMOS punchthrough stop (*n*-type); E, NMOS channel stop (*p*-type); F, PMOS channel stop (*n*-type); G, Retrograde *p*-well; H, Retrograde *n*-well; J, NMOS source/drain extension (*n*-type); K, PMOS source/drain extension (*p*-type); L, NMOS source/drain (*n*-type); M, PMOS source/drain (*p*-type); N, Triple well (*n*-type).

capable of damage gettering. An implanted buried layer is more effective than a highly doped epi substrate in suppressing latch-up [61], because it is located closer to the active devices. The position of a buried layer is set by the highly reproducible implant energy, and need not be set excessively deep to prevent well counterdoping by the epi substrate dopant. Buried layers can also be formed with non-dopant species, providing gettering benefits but no latch-up suppression.

A successful buried layer process requires sufficient implant dose and proper thermal processing to engineer the buried layer damage into stable dislocation loops at EOR [56,58,62]. Improper or excessive annealing can result in leakage-inducing line dislocations propagating to the near surface regions or counterdoping of the well regions by the buried layer dopant. When compared to channel and well implants, buried layers are relatively insensitive to dose precision or shadowing effects. The high doses of buried layers require that the implanted beam be energy pure to prevent counterdoping of the wells.

7.3.3 Source/Drain Implants

These implants are used to form highly doped ($>10^{20}$ cm^3) source and drain regions adjacent to the more lightly doped ($\sim 10^{18}$ cm^3) active channel and well regions of the device (Figure 7.23). The Source/Drain (S/D) regions are doped with the opposite conductivity type from the surrounding well to create diode isolation. Reverse-biasing of the drain junction is typical. S/D structures must be heavily doped to minimize MOSFET parasitic resistance and contact resistance to the silicide (see chapter on silicidation). To suppress punch through and the dependence of V_T on channel length (short-channel effect), the S/D junction must be very shallow, and become even more shallow as gate widths decrease.

Shallow S/D junctions for *n*-channel metal oxide silicon field effect transistors (NMOSFET) are usually formed by arsenic implantation. Shallow *p*-type junctions are more difficult to form due to the low mass

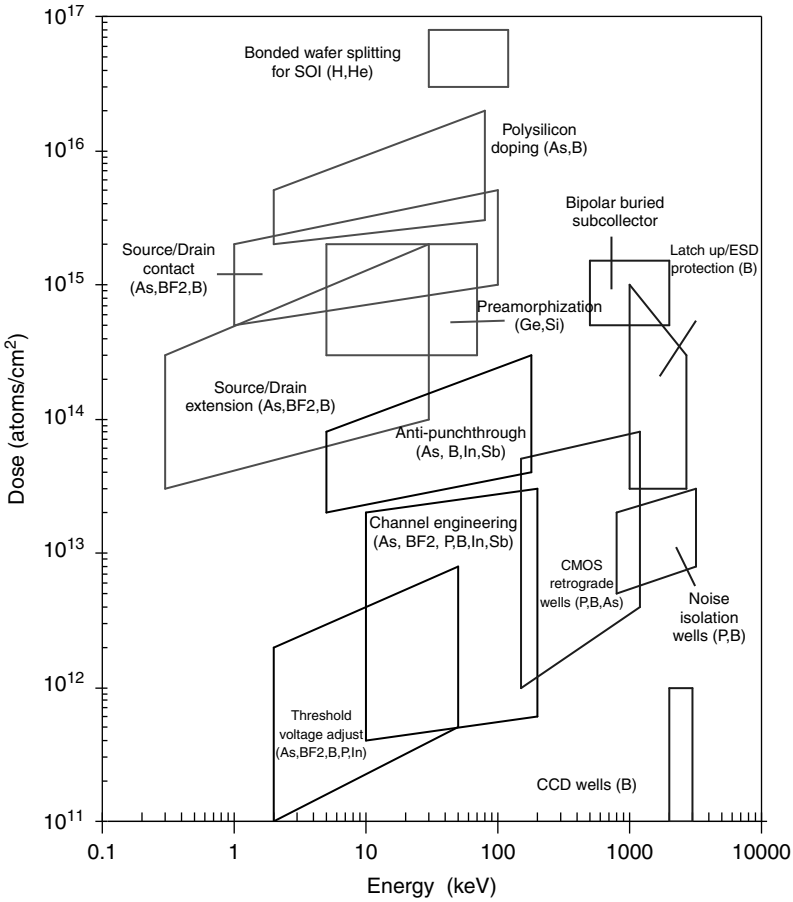


FIGURE 7.24 Typical dose and energy ranges for the doping applications in Figure 7.23. Within the energy range for each application, lower mass species are typical implanted at lower energies than heavier species. The dose of the application depends on device design requirements and is generally independent of species.

and susceptibility to TED of boron, the only *p*-type dopant with solid solubility suitable for this application. This has driven boron implant energies to the ultra-low (0.5–5 keV) range. Channeling tails are insignificant compared to TED effects below 5 keV, making optimization of the post-annealing very important. Instead of low energy boron, BF_2 can be implanted at higher energies. The S/D implants are usually done at 0° tilt or in a quad configuration to avoid device asymmetries. Beam energy purity is of paramount importance for S/D formation. Any high energy contaminants in the beam (e.g., neutrals arising from problems with deceleration mode) will result in deeper junctions and degraded devices. Low implant energies also preclude the use of screen oxides to capture sputtered-on metals, requiring minimal metallic contamination from the implanter. The dose rate (beam current) reproducibility is also important, since *p*-type S/D implants are usually close to the silicon amorphization threshold.

7.3.4 Source/Drain Extension Implants

The S/D extension (SDE) is a highly doped ($\sim 10^{20}$ cm^3) region under the sidewall spacer between the S/D regions and the channel (Figure 7.23), where lightly doped drain (LDD) structures [63] used to be. The LDD structure forced a trade-off between hot carrier immunity and transistor drive current [64–66].

As power supply voltages were reduced from 5 V to as low as 1.0 V, the hot carrier problem became less severe, while the parasitic resistance contribution increased. To offset this, the doping has been increased from the range of $1–5\times10^{13}$ to the $5–10\times10^{14}$ cm^{-2} making this region a low resistivity extension of the S/D structure. Because it is directly adjacent to the channel, a low SDE junction depth is even more crucial for punch through resistance than is the S/D depth. Since there are no silicide contact issues in the extension region, the dose and energy of this implant are both lower than for the adjacent S/D implant. Implemented properly, an aggressively scaled SDE junction can eliminate the need for a high tilt punch through stop implant, simplifying the process and the channel architecture. Phosphorus and arsenic are used for *n*-channel metal oxide silicon (NMOS) SDE implants, while boron is used for *P*-channel metal oxide semiconductor (transistor) PMOS devices. Implanter requirements for extensions are similar to S/D structures, energy purity, and low metallic contamination are required.

Control of the gate to extension overlap length to within approximately 5 nm is crucial for optimal device performance. Inadequate overlap degrades drive current and increases short channel effects, while extra overlap increases parasitic capacitance [67,68]; device speed is limited in either case. Proper positioning of the extension is achieved by implanting at 0° tilt or quad repositioning to ensure device symmetry. However, certain flash memory designs use an uncompensated 7° tilt for this implant; the resulting asymmetry improves programming and erasing efficiency [69].

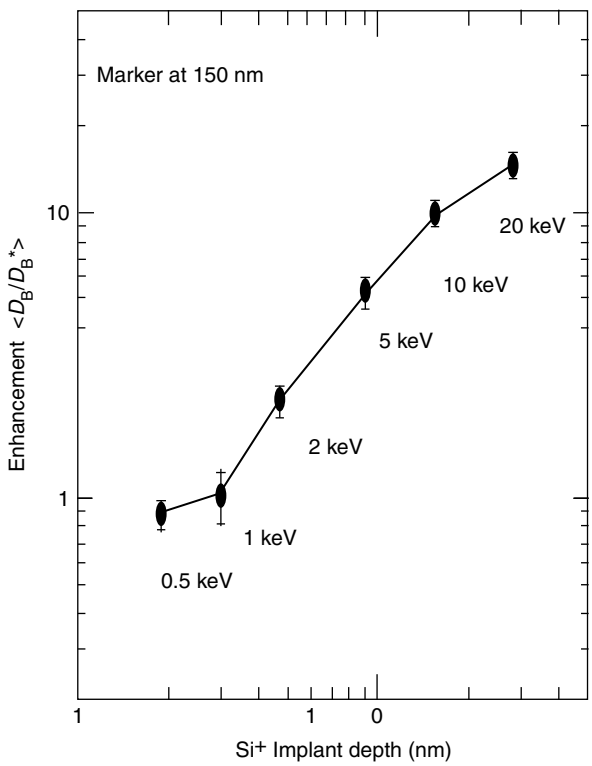


FIGURE 7.25 Enhancement in diffusion of a boron marker layer, grown by molecular beam epitaxy during a 950°C/30 s anneal, following implantation of 1×10^{14} cm^{-2} Si at various energies. (From Agarwal, A., et al., *IEDM Tech. Dig.*, 367, 1997; Gossmann, H. J., Rafferty, C. S. and Keys, P., *Materials Research Society Symposium Proceedings*, 610, B1.2.1, 2000.)

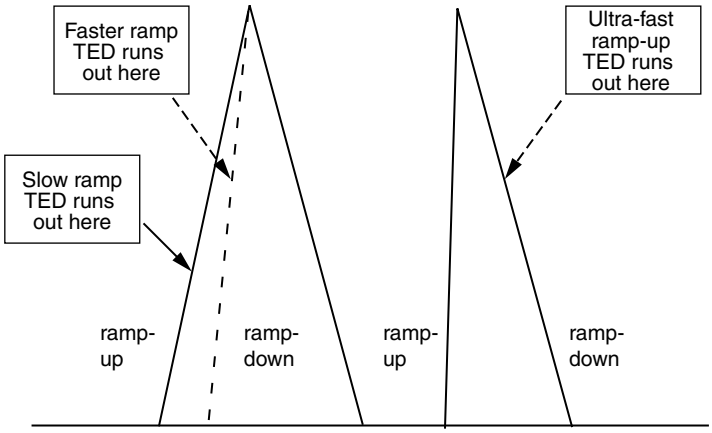


FIGURE 7.26 Schematic illustration of transient enhanced diffusion (TED) continuing during ramp down of a spike anneal that is sufficiently fast. (From Agarwal, A., Gossmann, H.-J.L., Fiory, A.T., Venezia, V.C., and Jacobson, D.C., *ECS Proc.*, 2000-9, 2000.)

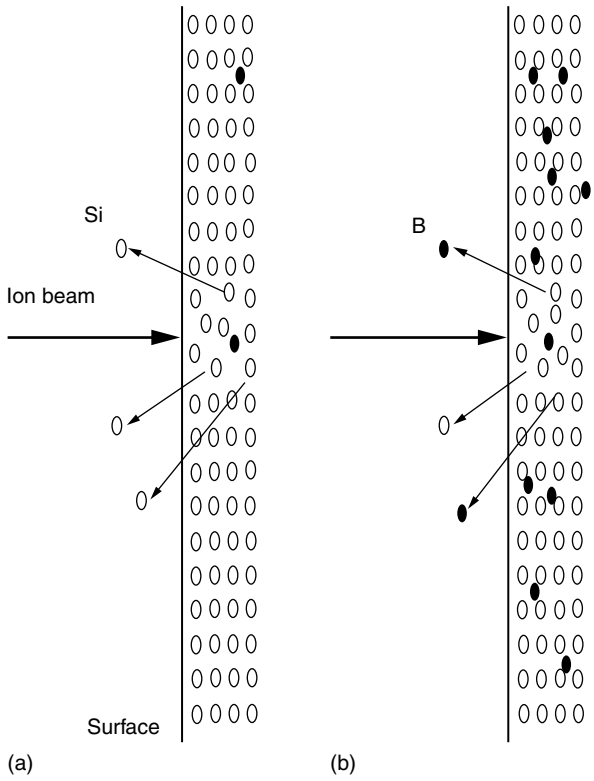


FIGURE 7.27 The sputtered dose depends on how much dopant has already been implanted. (a) At the start of the implant most of the sputtered atoms are silicon. (b) As the boron fraction in the near surface layers increases, more boron atoms are sputtered. (From Agarwal, A., Gossmann, H.-J. L., Fiory, A.T., Venezia, V.C., and Jacobson, D.C., *ECS Proc.*, 2000-9, 2000.)

7.3.5 Ultra-Shallow Junction Formation for Source/Drain Extensions

Present device-scaling rules demand that the vertical and the lateral dimensions be reduced by a factor of two, approximately every 5–6 years [70]. Accordingly, both the deep S/D and the SDE junction depths need to be reduced by the same factor [70]. The main challenge in junction scaling is meeting the junction depth requirement while maximizing the active dopant concentration. It is also important to minimize the electrically active residual damage, which increases the junction reverse leakage current and thus the transistor off-state leakage current. Electrically active defects are of concern only if they are located in the junction depletion region or within the carrier diffusion length of the depletion region. It has been known for some time that boron diffusion can be enhanced by damage introduced by the implant process. For example, Figure 7.25 shows the enhanced diffusion of a boron marker produced by molecular beam epitaxy on a silicon substrate, which was subsequently damaged by 1×10^{14} cm^{-2} silicon implants at various energies and then subjected to a 950°C/30 s anneal. The enhancement scales linearly with the projected range of the implant which is approximately where the damage induced excess interstitials are initially located [71,72].

The phenomenon of TED after ion implantation increases the challenge of forming ultra-shallow junctions [73–75]. Ion implantation leads to the displacement of silicon atoms from their lattice positions, creating pairs of vacancies and interstitials. During the initial stage of post-implantation annealing most of the vacancies and interstitials recombine leaving behind a net excess of interstitials

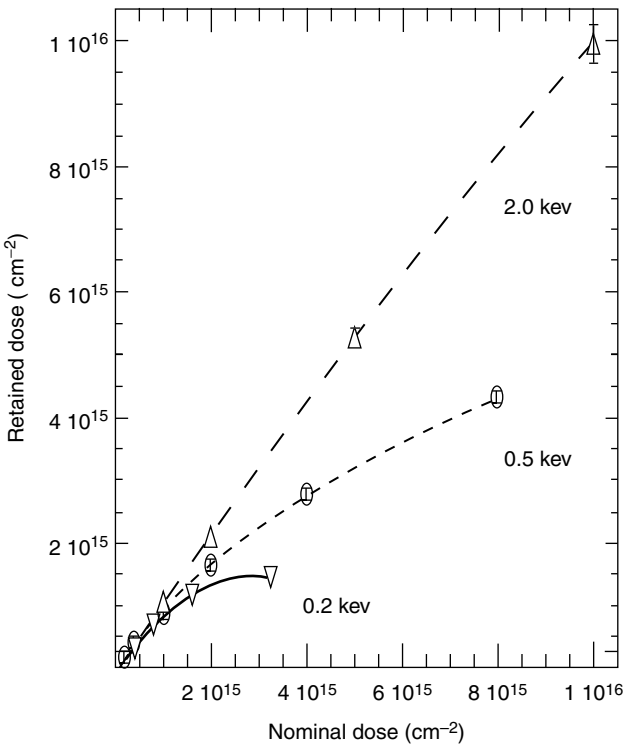


FIGURE 7.28 Retained B dose (measured using the B(p,α)Be nuclear reaction) vs. nominal B dose, implanted into (100) silicon wafers with an as-received native oxide. The trend lines are 3rd order polynomial splinefits. For a nominal B dose of 1×10^{15} cm^{-2}, sputtering leads to a ~10% reduction in the retained dose at 0.5 keV, or a ~20% reduction at 0.2 keV. (From Agarwal, A., Gossmann, H.-J.L., Fiory, A.T., Venezia, V.C., and Jacobson, D.C., *ECS Proc.*, 2000-9, 2000; Agarwal, A., *2000 Conference on Ion Implantation Technology Proceedings*, ed. Ryssel, H., Frey, L., Gyulai, J., and Glawischnig, H., IEEE Press, Piscataway, NJ, 293, 2000.)

approximately equal to the implanted ion dose; this is also referred to as the "+1" approximation [76]. These excess interstitials quickly coalesce into extended defects, such as {311}'s [77,78], or more stable dislocation loops. While these extended defects have lower free energy than individual interstitials [78,79], they are still metastable and dissolve with continued annealing. As they dissolve, they release excess interstitials into the lattice. Since boron diffuses by an interstitial mechanism [80] its diffusivity is enhanced by the excess interstitials with the time averaged diffusivity enhancement equal to the time averaged interstitial supersaturation [80]. Both the interstitial supersaturation and the diffusivity enhancement end soon after the defects have dissolved. The increase in junction depth, Δx_j, due to TED to be expressed as [79,81]

$$\Delta x_j^2 \propto N\, R_p \exp[-(-1.4\text{ eV})/kT] \tag{7.36}$$

where N is the number of interstitials trapped in the defects (approximately equal to the implanted dose) and R_p is the projected ion range (where the excess interstitials are initially located). The linear dependence on R_p has been demonstrated experimentally (Figure 7.25) [71,74]. The activation energy of Δx_j^2 is negative because the interstitial supersaturation due to the presence of the extended defects is larger at lower temperatures. This implies that the final junction will be deeper if the defects are annealed out at a lower temperature than at a higher temperature. This is a key reason why junction anneals are done in a rapid thermal anneal (RTA) rather than in a conventional furnace with a ramp-up rate of a few degrees per minute. An RTA spends significantly less time during the temperature ramp-up at lower temperatures where the diffusivity enhancement is larger.

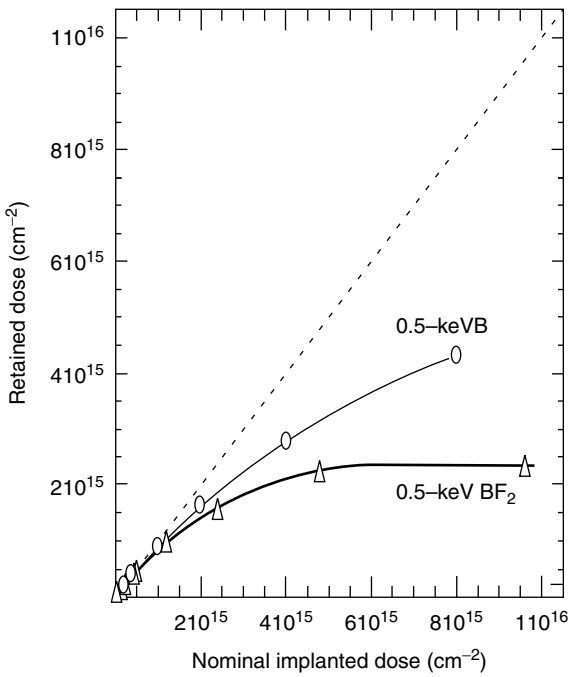


FIGURE 7.29 Retained B dose (measured using the B(p,α)Be nuclear reaction) vs. nominal B dose in wafers implanted with either B or BF_2 at an equivalent boron depth. The trend lines are 3rd order polynomial splinefits. For a nominal dose of $1\times10^{15}\,cm^{-2}$, sputtering leads to a ~10% reduction in the retained dose for 0.5-keV B, or a ~20% reduction at 2.2-keV BF_2. (From Agarwal, A., Gossmann, H.-J.L., Fiory, A.T., Venezia, V.C., and Jacobson, D.C., *ECS Proc.*, 2000-9, 2000.)

Since the increase in junction depth due to TED depends on the implant dose (Equation 7.36), it is possible that for a high dose implant some damage will remain after a fast ramp-up, allowing TED to continue during the ramp down [75]. As the ramp-up rate is increased, the temperature at which TED runs out is pushed up until the TED is pushed over to the ramp-down side of the anneal (Figure 7.26) [82]. This is illustrated in Figure 7.26.

Sputtering of the target and dopant atoms during implantation can also cause boron dopant loss as shown in Figure 7.27 [82]. Sputtering of the target during implantation is an important concern with ultra-shallow implants. For a target atom to be sputtered it has to acquire enough energy to overcome the surface binding energy of the target [83]. Though sputtering occurs at any implant energy, the number of atoms sputtered per incoming ion increases as the implant energy decreases [83]. At an energy of 0.5 keV approximately one target Si atom is sputtered for every four B atoms implanted [83–85]. For energies low enough that a significant number of implanted atoms come to rest in the near surface layer, more and more boron atoms will be sputtered with increasing B dose (Figure 7.27 and Figure 7.28). Increased dopant sputtering with increasing dose causes the fraction of implanted dose that is retained to decrease with increasing dose. This effect is even more severe at energies less than 0.5 keV. This is very undesirable given the difficulties associated with achieving large beam currents at such low energies. Using BF_2, further exacerbates this problem, as the co-implanted F atoms

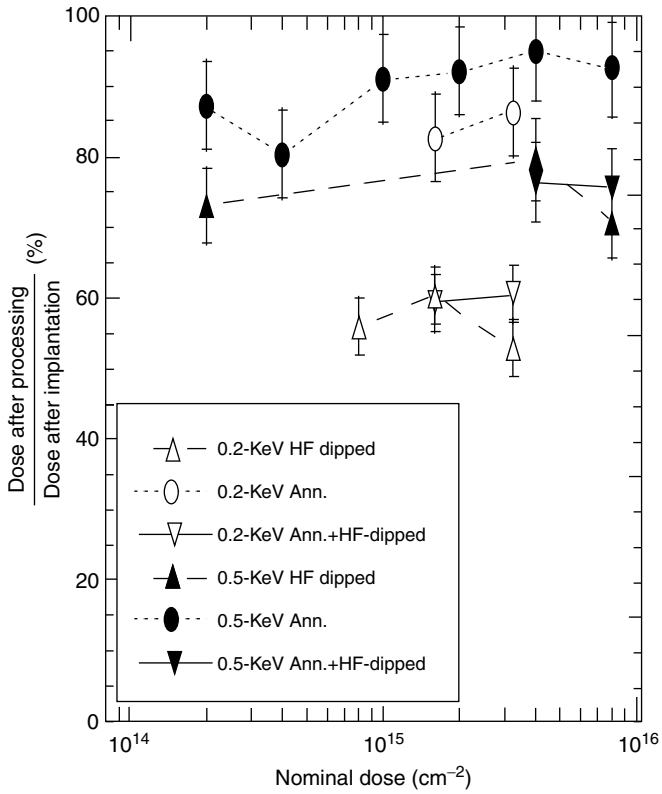


FIGURE 7.30 Fraction of B retained in wafers after oxide removal (60 s HF treatment), or after annealing in a 0.5% O_2 ambient for 15 s at 1000°C, or after annealing followed by oxide removal. Comparing the HF dipped data with the Annealed + HF dipped data shows that the dopant lost during annealing came from the oxide. (From Agarwal, A., Gossmann, H.-J.L., Fiory, A.T., Venezia, V.C., and Jacobson, D.C., *ECS Proc.*, 2000-9, 2000; Agarwal, A., *2000 Conference on Ion Implantation Technology Proceedings*, ed. H., Ryssel, L., Frey, J., Gyulai, and H., Glawischnig, IEEE Press, Piscataway, NJ, 293, 2000.)

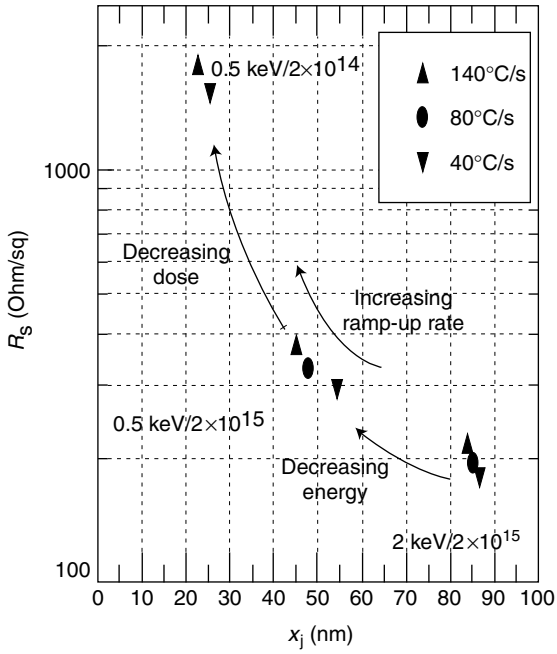


FIGURE 7.31 Sheet resistance vs. junction depth as a function of ramp rate, implantation dose and implantation energy. Note the similarity between increasing the ramp-up rate or reducing the energy and dose. (From Agarwal, A., Fiory, A.T., Gossmann, H.-J., Rafferty, C. S., Frisella, P., and Hebb, J., *Mater. Sci. Semicond. Process.*, 1, 237, 1999; Agarwal, A., Gossmann, H.-J. L., Fiory, A. T., Venezia, V.C., and Jacobson, D.C., *ECS Proc.*, 2000-9, 2000.)

contribute to the boron sputtering but not to the retained B dose (Figure 7.29). Additional dopant can be lost to oxide removal and annealing effects (Figure 7.30). These figures make it clear that caution must be exercised when comparing junction data from the literature for different sub-keV implants or for different species. A correction needs to be applied for the different retained doses. While a junction from a lower energy implant is indeed expected to be shallower because of reduced TED, so is a junction from a lower dose implant at the same energy.

In the sub-kiloelectron volt regime, there is more than one way to arrive at the same junction properties. It is very important to minimize the dose first, before reducing the energy further. The dependence of the sheet resistance and junction depth data on the different implant and annealing parameters is summarized in Figure 7.31. Increasing the ramp-up rate leads to a more shallow junction with higher resistivity. The same is also true when a smaller dose or energy is used. Modifying the implant parameters first helps to avoid the risk of poor process repeatability which necessarily accompanies the use of higher ramp-up rates.

7.4 Commercial Ion Implantation Equipment

7.4.1 Introduction

The concept of doping by ion implantation to change semiconductor conductivity was disclosed in the original 1954, patent by Shockley [86]. This patent details much of the implantation and annealing art as it is featured today. Implantation, as a doping technology, started to displace conventional diffusion techniques in the early 1970s with the emergence of MOS transistors and the recognition that the achievement of reproducible MOSFET threshold voltage shift could only be achieved by implantation.

According to a recent review by McKenna [87] the first true commercial ion implanters were manufactured in the early 1970s, (Accelerators Inc. 200 MP and Extrion Inc. 200–20). These early implanters were typically capable of producing a few microamperes of beam current with energies of up to 200 keV. In 1978, Nova (now Axcelis Technologies) introduced the first true high current ion implanter, the NV 10–80. This implanter was capable of producing 10 mA of beam current at energies up to 80 keV. Today, there are over 4000 ion implanters in IC factories throughout the world.

Complex implanted dopant structures are required for modern integrated circuits. For example, deep buried layers for device isolation can extend to depths of approximately 1 μm requiring the use of MeV implantation, whereas MOSFET drain extension implants require sub-keV energies and halo implants require implantation into tilted wafers ($\sim$45°). Figure 7.24 of Section 7.3 gives a broad perspective of the breadth of the dose-energy requirements for modern integrated circuits. This broad diversity of demands has led to the natural evolution of three classes of machine: (1) high Current Implanters ($\sim$500 to $\sim$100 keV), (2) medium Current Implanters ($\sim$1 to $\sim$300 keV) with high tilt capability, and (3) High Energy (also called High Voltage) Implanters ($\sim$10 keV to $\sim$2 MeV). Within these guidelines one can construct a general specification (Table 7.11) for the suite of nominally 20–30 implantation machines in a typical leading edge 200 or 300 mm Fab.

Of that the particular importance are:

- the capability to implanting a wide variety of species such as shown in Table 7.11;
- dose uniformity requirements, across the wafer, wafer to wafer, and lot to lot;
- particulate, metal and cross-species contamination requirements;
- productivity requirements, in both small lot and large lot environments.

Central to these productivity requirements is ion source performance where glitch-free operation, maximum achievable beam currents and short recipe change tune times directly impact wafer throughput and set up times. For example, present generation implanters require ion source tune-times of less than 3 min for implant recipe that changes to involve new selections of species, energy, and dose. Also crucial to productivity is the wafer transport system that must achieve rapid transport of wafers in an out of the evacuated implantation chamber and also minimizes the time when the implant chamber is in the idle state awaiting the availability of the next wafer. In addition important to productivity is the means used to translate the ion beam with respect to the wafer in such a way that

TABLE 7.11 General Requirements for Present Generation Ion Implant Equipment

Dose and Energy Capabilities	
Dopant species	B, BF_2, P, As, In, Sb, $B_{10}H_{22}$, $B_{18}H_{22}$
Other species	H, He, N, O, Si, Ar, Ge, Xe, C
Energy range	0.250 keV–3 MeV
Beam current range	0.5–30 mA
Dose range	1×10^{11}–1×10^{16}/cm^2
Implant Process Capabilities	
Dose uniformity	$<$0.5%
Energy integrity	$<$1.0%
Implant angle range	0–60°
Implant angle integrity	$<$1.0°
Elemental contamination	$<$10 ppm of implanted dose
Partical contamination	$<$0.10/cm^2 for particles $>$0.12 m
Equipment Productivity	
Mechanical limit wafer throughput	Up to 400 wafers/h
Ion source lifetime	100–1000 h
Energy and species change tune time	$<$3 min
Mean time between failure	$>$100 h
Total availability	$>$85%

maximizes the utilization of the available ion beam, while still enabling the achievement of dose uniformity, implant angle control, and implant angle integrity across the wafer. Finally, the system must enable the cooling of the wafer such that photoresist masks may be used to control the implanted regions on the semiconductor chip. The engineering solutions and resulting system architectures will be summarized in this section.

7.4.2 Implanter Architecture Drivers

The function of an ion implanter is to efficiently produce a beam of selected ions that is pure in species, energy, and charge state, and to uniformly and productively implant these over the entire surface of the wafer. Lateral control of implantation into the integrated circuit chip is achieved by means of a photoresist mask, or the use of a device feature such as to gate stack. Given these requirements, all implanters have the following components:

- An ion source capable of producing abundant quantities of ions having the desired species and charge state.
- Extraction optics that extract the ions from the source, accelerate these to some predetermined energy and shape the resulting ion beam.
- A mass analysis system that removes unwanted species from the ion beam.
- Additional accelerating/decelerating optics (if needed) that regulates the final energy of the extracted beam that impacts on the wafer.
- A beam/wafer scanning system that enables the uniform painting of the ion beam across the wafer(s) surface.
- A dose measurement and control (dosimetry) system.
- A charge neutralization system that manages the charging of the wafer that is exposed to the ion beam.
- A wafer handling system that transports wafers in and out of the target chamber pedestal where wafer cooling is provided.
- A vacuum system to control the low pressure ambient required in the ion transport and wafer target chambers.

7.4.3 Ion Sources and Extraction Optics

Regardless of system architecture, the ion source is one of the most important ion implanter subsystems because it dictates not only the process capability of the system, but also plays an important role in system throughput and set up time. Great demand is placed on the ion source because, unlike most other accelerators where the ion source is required to produce only a single ion species, the ion implanter requires an ion source capable of the contamination-free, on demand production of the variety of different species shown in Table 7.11. The difficulty of this task has historically limited the suitable source types to the Freeman and Bernas sources. Both these sources use hot cathodes that emit ionizing electrons to produce a low voltage arc discharge. Both sources also use an axial magnetic field to confine the ionizing electrons to enhance the degree of discharge ionization. A more comprehensive overview of the physics of electrical discharges used in ion sources is given in [88]. The Freeman ion source [89] was the workhorse of earlier implanters; however limited source life and available beam currents have resulted in its replacement by the enhanced Bernas ion source [90]. All present day ion implanters use some version of Bernas type ion source.

7.4.3.1 Bernas Source and Extraction Optics

Figure 7.32 schematically depicts the Bernas source and extraction optics. The source consists of an arc chamber usually fabricated from a refractory metal, such as tungsten or molybdenum. Interior to the chamber is a resistively heated cathode, biased negatively with respect to the arc chamber. Electrons emitted

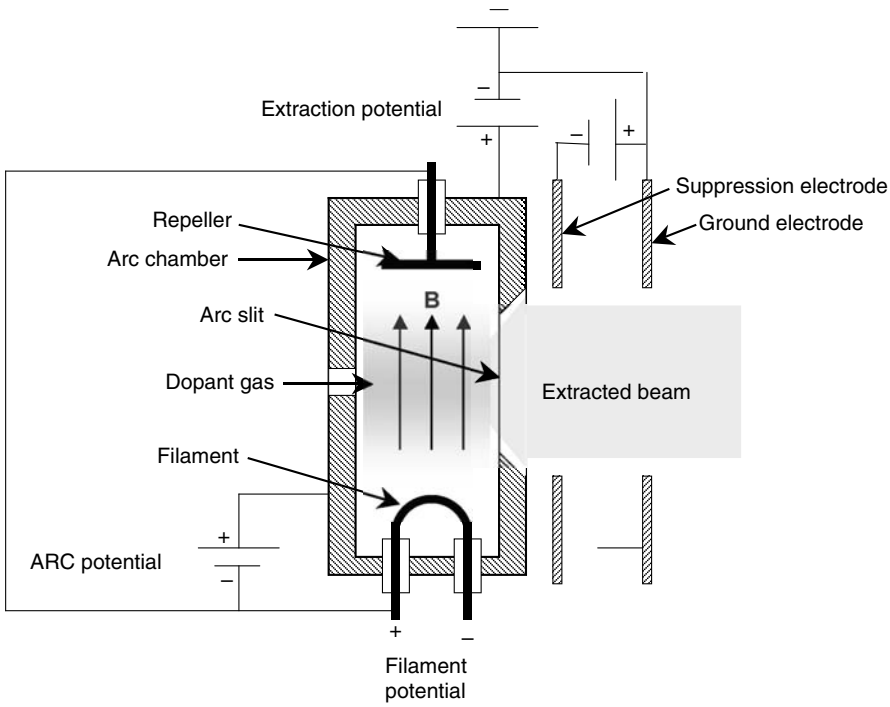


FIGURE 7.32 Enhanced bernas ion source with extraction optics.

by the source are accelerated by this bias potential and are launched into the arc chamber causing ionization of the dopant gas that is introduced into the arc chamber through a small orifice. The applied magnetic field (designated by B in Figure 7.32) causes the ionizing electrons to helically drift in the direction of the magnetic field toward a negatively biased repeller electrode. Upon approaching the repeller electrode, the drifting electrons reverse course and become redirected toward the negatively biased cathode where they are again repelled back into the arc plasma. This to- and fro-drift motion traps the electrons (Penning trap) in the source plasma with the result that very efficient ionization of the source gas ensues. As seen from the Figure 7.32, the ions are extracted in a direction that is perpendicular to the applied magnetic field. Other sources known as penning ion gauge (PIG) sources make use of a similar trap but differ in that ions are extracted in a direction that is parallel to the applied magnetic field [88].

Vaporized or gaseous feed materials are injected into the arc chamber to provide for the generation of the desired ion species. Table 7.12 gives a listing of the source materials that are most commonly used. These consist of either gaseous sources such as the hydrides or halides, or consist of directly vaporized solids (Arsenic, Phosphorus, or Antimony). Direct vaporization requires that the source be equipped with a small heated chamber and transfer line to inject the elemental vapor directly into the source. Source gases are provided as cylinders of the compressed gas, or as cylinders within which the gas is adsorbed onto an inert adsorption medium that allows release of the gas upon exposure of the

TABLE 7.12 Common Source Materials for Different Implant Species

Species	Gas Source	Solid Source	Species	Gas Source	Solid Source
Arsenic	AsH_3	Elemental Arsenic	Antimony	—	Sb_2O_3, Sb metal
Phosphorus	PH_3	Elemental Phosphorus	Indium	—	$InCl_3$
Boron	BF_3	—	Germanium	GeF_4	—
BF_2	BF_3	—	Silicon	SiF_4	—

vacuum-pumped source chamber [91]. The latter option is preferred since most source gases are highly toxic and/or flammable and therefore the compressed gas cylinders present a safety hazard.

As depicted in Figure 7.32 the resistively heated cathode of the Bernas source is directly exposed to the arc discharge where physical and/or chemical erosion act to limit cathode life. These problems may be overcome by indirectly heating a more robust cathode by some auxiliary means. An example of this is the Axcelis extended life source (ELS) [92]. This source (Figure 7.33) is based upon the design of the enhanced Bernas source [90], but the filament-style cathode is replaced by a robust cathode structure that is indirectly heated by electron bombardment from another thermionic negatively biased (~1000 V) filament. This filament however is isolated from the source plasma, and therefore protected against erosion. With this configuration, the arc cathode may be made of a much heavier erosion-resistant design. References 93–95 provide some performance data for this type of source. Further enhancements to this source design have been achieved by replacing the repeller with a second indirectly heated cathode structure [96]. This latter source design has yielded enhanced beam currents and extended lifetimes especially for multiply charged ions. Reference 97 provides a valuable compilation of procedures and tips related to operation and maintenance of the Freeman, Bernas, and Indirectly Heated Cathode Bernas sources.

Referring again to Figure 7.32, it is noted that the ion source arc chamber is biased positively with respect to ground by the variable extraction potential, which can be as high as 100 kV. A ground electrode slit, positioned adjacent to a similar slit in the arc chamber serves to extract and accelerate ions from the source and launch these into beamline. From Figure 7.32 it is noted that there is a second slit electrode positioned between arc chamber and ground electrode named the suppression electrode. This negatively biased electrode serves as a barrier to low energy electrons present in copious quantities in the beamline upstream of the extraction electrode and thereby prevents these from being accelerated into the positively based arc chamber. The suppression electrode serves two important purposes. Firstly, by preventing the extraction of low energy electrons from the beamline it helps to preserve space charge

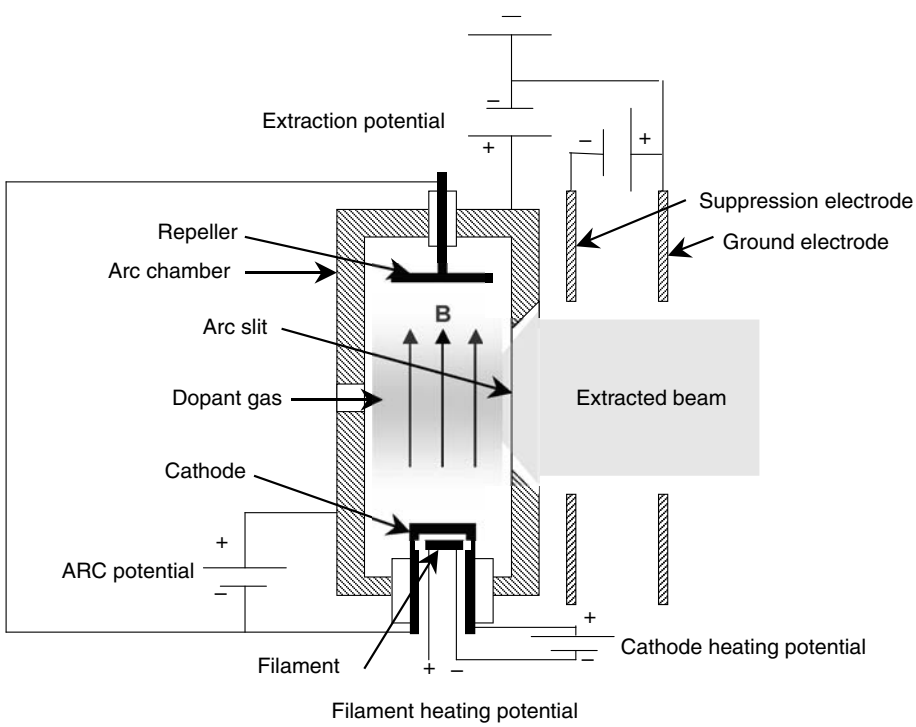


FIGURE 7.33 Enhanced bernas source with indirectly heated cathode.

neutrality in the ion beam thereby preventing beam blow up. Secondly, by preventing electron acceleration into the source, the suppression electrode minimizes x-ray generation, and the large extraction currents that would otherwise ensue. The gaps between the electrodes can be varied in situ and the suppression electrode can be tilted and/or can have a transverse motion to properly steer the extracted beam into the mass resolving magnet.

The extraction potential determines the ion energy in the beam. Therefore, for ion implansters that use no other means of increasing or decreasing ion energy, the extraction potential is the sole determinant of ion energy. In present day, high current ion implanters this is often the case and the extraction system must therefore provide for the generation and transport of ion beams with energies as low as 500 eV (for ultra-shallow MOSFET S/D junctions) up to the approximately 100 keV. This challenge has resulted in the development of more complex extraction electrode [98] systems that use higher extraction potentials to generate higher ion currents and subsequently decelerate the ions prior to injection into the mass resolving magnet.

7.4.3.2 Molecular Ion Sources

In recent years, significant advances have been made in the development of molecular beam sources for dopant implantations into silicon. The driver for the development of these sources has been the need for very low energy implants. Energy is partitioned between the atoms of a molecule in direct proportion to their mass. For example, the widely used molecular ion BF_2^+ with atomic mass $\sim$49 having a single boron atom of mass approximately 11 results in the implantation of boron at an energy that is approximately 11/49 of the molecular ion energy, for example, a 10 keV BF_2 implant is energetically equivalent to a 2.24 keV B implant. Similarly, dimers of arsenic [99] and phosphorus (As_2^+ and P_2^+) have been used to facilitate the formation of shallow arsenic or phosphorus junctions. Here, there is a 2:1 energy partition and equally important the dose rate is twice the beam current. These ions may all be generated in the previously described Barnas sources.

A much more dramatic example of this energy partitioning may be achieved with decaborane ($B_{10}H_{14}$) [100] where 10 keV implant is equivalent to approximately 1 keV implant. Recently, another large boron containing molecule, Octadecaborane ($B_{18}H_{22}$) has also been identified as a useful molecule for this application [101]. It is important to note that with these molecules, 1 mA ion beam current is equivalent to 10 (for decaborane) or 18 mA (for octadecaborane) of boron current. For this reason the molecular beam obviates many of the space charge limitations associated with the ultra-low energy Boron beams. Conventional Bernas or Freeman ion sources are not suitable for decaborane or octadecaborane implantation since the high arc chamber temperature causes disassociation of the molecule. Ionization chamber temperatures below 300°C and a different approach to electron impact ionization of the

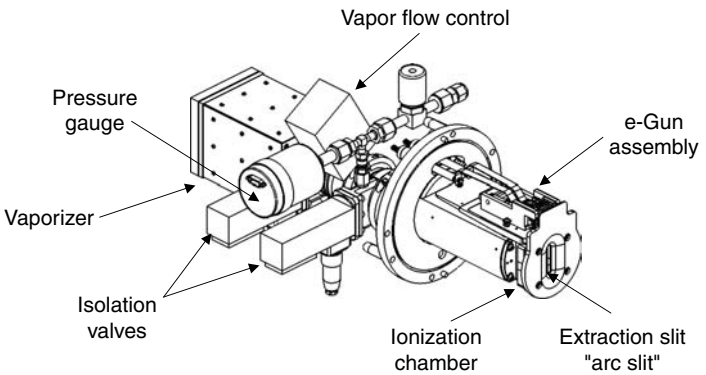


FIGURE 7.34 Ion source suitable of decaborane or octadecaborane ion beam generation. (From Jacobson, D. C., Private Communication, SemEquip Co., 2005.)

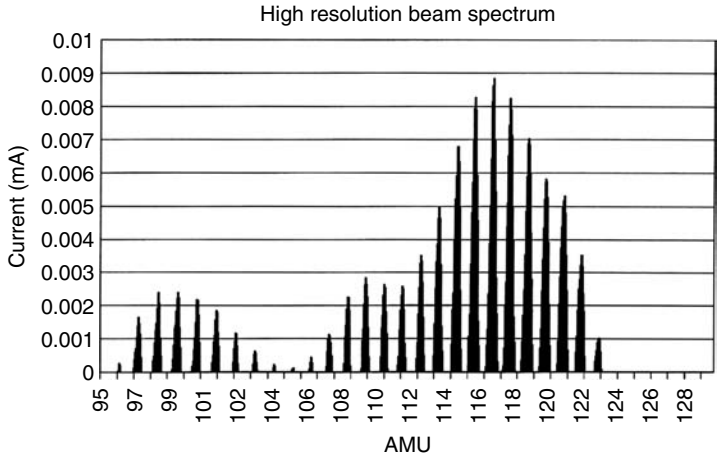


FIGURE 7.35 Typical octadecaborane spectrum.

molecule are required [101]. Figure 7.34 shows a commercially available octadecaborane ion source [102]. In addition the ionization process results in a distribution of ions of the form $B_{10}H_x$ or $B_{18}H_x$ with the result that the mass resolved spectrum consists of a typically up to 10 peaks, all containing the same boron content but with varying hydrogen content. As a result, the acceptance of the mass resolving system must be increased to allow for maximum utilization of the available molecular ion current [101]. Figure 7.35 gives a typical mass resolved spectrum obtained from a decaborane source [102].

7.4.4 Mass Analysis Systems

7.4.4.1 Dipole Magnet Mass Resolving Systems

Ion sources of the types just discussed above produce a variety of ion species with different charge states. These species are all extracted and injected into the mass analysis system. Magnetic analysis is used in all commercial ion implanters to sort out the desired dopant ions from ion population extracted from the source. Figure 7.36 depicts a typical ion source and mass resolving system. Depicted here is a source at

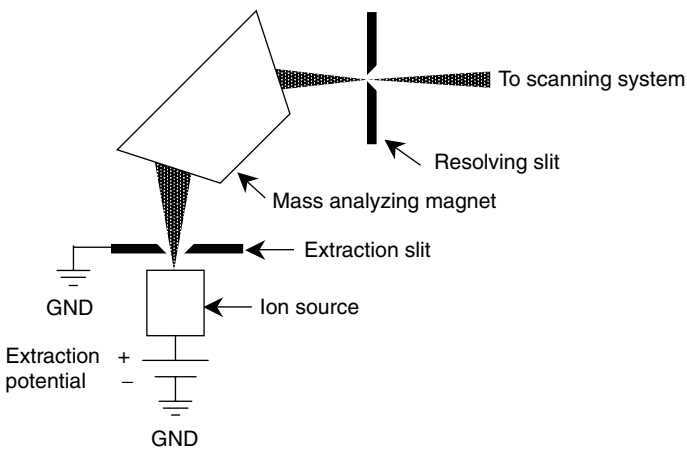


FIGURE 7.36 Magnetic ion mass resolving system.

elevated positive potential with the extraction grid system at ground (the suppression electrode is not shown) that directs a beam of ions into a mass analyzing magnet. The object in Figure 7.36 labeled "Mass Analyzing Magnet" represents two pole pieces of an electromagnet. These pole pieces lie above and below the plane of the figure, with ion beam (shown as the shaded object) passing between. The magnetic field that passes from one pole piece to the other lies normal to the plane of the figure. The long axis of the arc slit of the ion source typically lies parallel to the magnetic field and therefore also lies perpendicular to the plane of Figure 7.36. The ion beam is dispersed by the field and for this reason the plane of Figure 7.36 is called the dispersion plane of the ion implanter. The orthogonal plane (usually parallel to the long axis of the ion source slit) is named the non-dispersive plane. As will be shown in the following, ions having a velocity component parallel to the magnetic field do not experience a magnetic force. Therefore the parallel velocity component is unaffected.

The magnetic force on a beam of moving ions is given by the following vector equation:

$$\vec{F} = q\vec{v} \times \vec{B} \tag{7.37}$$

where $\vec{F}$ is the magnetic force, q the charge on the ion, $\vec{v}$ its velocity vector, and $\vec{B}$ the magnetic field vector. The symbol $\times$ refers to the vector cross product. This equation implies that the component of the ion velocity $v_\perp$ that is normal to the magnetic field direction will experience a force $F = qv_\perp B$, whereas the velocity component parallel to the magnetic field will remain unaffected. Referring again to Figure 7.36, it is noted that the ion velocity lies in the plane of the figure, and the magnetic field normal to the plane of the figure. We therefore dispense with the term $v_\perp$ and use the term v in the following analysis. Equation 7.37 also dictates that the magnetic force is mutually perpendicular to the velocity and magnetic field directions which would tend to force the ion into a circular orbit, the radius of which is determined by equality between the centripetal magnetic force and the inertial centrifugal force. The motion is therefore determined by Equation 7.38 below,

$$\frac{mv^2}{R} = qvB \tag{7.38}$$

where R is the radius of curvature of the ion trajectory. In addition, the ion velocity is related to the energy imparted by the extraction potential by Equation 7.39 below.

$$E = \frac{1}{2}mv^2 \tag{7.39}$$

Therefore according to Equation 7.38 and Equation 7.39, the radius of the ion trajectory is uniquely determined by the ion charge, mass, energy, and magnetic field intensity. A resolving slit at the exit of the analyzing magnet will block all species except those having the proper trajectory radius R for a given magnetic field intensity B. The product, RB, called the magnetic rigidity, quantitatively defines these trajectory requirements and may be derived from Equation 7.38 and Equation 7.39 above:

$$RB = \sqrt{\frac{2mE}{q^2}} \tag{7.40}$$

In practice, the magnetic field generated by the mass resolving magnet is adjusted until the desired ion species travels along that trajectory radius that allows its transport through the mass resolving slit, as depicted in Figure 7.36. In addition, as also depicted in Figure 7.36, the pole pieces are shaped in a manner that causes the diverging ion beam entering the magnet to be focused at the resolving slit. A detector placed at the site of the resolving slit would then typically show an ion distribution having a Gaussian-like profile, having a peak at the radius position R, corresponding to an ion mass M and some half-width, equivalent to a mass range ΔM. Within the limits of M and ΔM the mass resolving slit serves

to block the transport of all ion beams except the beam of the desired specie which is allowed to pass and travel on to the wafer implantation site. Mass resolution, defined as center mass M, in AMU for a particular peak divided by its full width at half maximum, ΔM, is a very important figure of merit in any mass spectrometer system. It indicates the degree to which the system is capable of separating beams of different species with nearly the same beam rigidity. If the trajectories of two beams incident on the resolving slip are separated by a gap that is larger than mass of the smaller ion divided by ΔM then the two beams will effectively be separated, and high species purity achieved. In commercial ion implanters the mass resolution can vary from as low as 10 to as high as 100.

7.4.4.2 Species and Energy Contamination in Magnetic Mass Resolution Systems

Two forms of species contamination may be transmitted in a magnetic mass resolution system. The first, called mass interference, results when two or more different chemical species that have nearly the same mass are transported through the mass resolving magnet but are not resolvable within the $M/\Delta M$ limits of the mass analysis system. Important examples of mass interference are discussed in Section 7.5.7 of this chapter. The second variety of contamination arises when circumstances permit the generation of two ions of the same chemical species having different energies but the same magnetic rigidity. These result in an energy contamination and dose error that is particularly important when implanting with multiply charged ion species, e.g., As^{++} or with dimers, e.g., As_2^+. This contamination type is important with arsenic and phosphorus implantation since the ion source produces a mixture of multiple charged ions (up to $3+$ state), as well as molecular species such as dimers, trimers, etc. Referring again to Equation 7.40 it may be determined that a doubly charged ion having an energy E, has the same magnetic rigidity as a singly charged ion of the same specie but with one-fourth of the energy. Similarly, a singly charged dimer has the same magnetic rigidity as that of a singly charged monomer ion having twice its energy. The former case leads to double-hump implant profile with an undesired low depth peak. The latter leads to a double-hump implant profile having an undesired high depth peak. These are discussed more, fully in Section 7.5.2 of this chapter.

7.4.4.3 Mass Resolving Systems That Produce Ribbon Beams

As previously mentioned, the shape of the magnet poles may be used to provide focusing of the ion beam in the dispersion plane of the mass resolving system. This concept has been used as a means for generating a ribbon-shaped ion beam [103]. To achieve this, a Bernas source is equipped with a convex-shaped arc slit, with matching convex extraction and suppression electrodes. Here, the long axis of the arc slit lies parallel to the dispersion plane. This configuration results in a diverging beam that is focused by the shaped magnet pole pieces onto the resolving slit. The beam diverges again upon passing through the resolving slit where added magnetic optics produce a uniform ribbon-shaped beam having a width that is greater than the wafer diameter, and having a "thickness" of several centimeters. Scanning is accomplished by mechanically translating the wafer in a direction perpendicular to the ribbon width. Additional magnetic optical elements are also required to assure uniform flux density over the entire ribbon width.

7.4.4.4 Transport of High Perveance Beams

In recent years, MOSFET device scaling has resulted in the need for very shallow high dose p–n and n–p junctions. This has resulted in the need to transport low energy, high current ion beams through the mass resolving system, with energies as low as 500 eV under investigation. Such beams are highly vulnerable to space charge blow up since for any given beam current and beam cross-sectional area, the number density of ions in the beam varies inversely with the ion velocity, $\sqrt{2E/m}$. As a result, low energy beams particularly of high mass ions such as arsenic, have a very high number density of ions in the beam, and unless space charge neutralization is provided and maintained the repulsive space charge blow up will prevent meaningful beam transport. The propensity for space charge blow up may be quantitatively

defined by the generalized beam perveance given by Equation 7.41 below [104]:

$$K = \frac{1}{4\pi\varepsilon_0\sqrt{2e}} \frac{\eta I\sqrt{m/Z}}{E^{3/2}} \tag{7.41}$$

here η is the fraction of beam current I, that is not neutralized, E is the ion energy, Z the ion charge number, e the electronic charge, and m the ion mass. Modeling work reported in [104] shows that a 2 keV B^+ 10 mA beam having 4% non-neutralized beam current expands by a factor of six in diameter after traveling 0.5 m due to purely space charge effects. The relationships between perveance and beam blow up is given in Equation 7.42a and Equation 7.42b below which quantify the space charge induced change in beam cross-section for a beam of ions traveling in the z-direction having transverse width $2a_x$ in the x-direction and height $2a_y$ in the y-direction.

$$\frac{\partial^2 a_x}{\partial z^2} = -\frac{2K}{a_x + a_y} \tag{7.42a}$$

$$\frac{\partial^2 a_y}{\partial z^2} = -\frac{2K}{a_x + a_y} \tag{7.42b}$$

Equation 7.42a and Equation 7.42b make clear that beam blow up can be minimized by making the beam cross-section very large, or by minimizing the perveance. The transport and implantation of large cross-section beams are usually impractical from the standpoint of equipment cost and beam utilization. Therefore commercial ion implantation equipment is engineered to minimize perveance. Clearly, without a high degree of neutralization, beams of high perveance cannot be transported through the mass resolving magnet, nor for that matter, through the remainder of the beamline. Some charge neutralization can arise from ion beam interactions with residual gas molecules in the beamline [105]. These interactions result in creation of electron-residual gas ion pairs in the beam path. The resultant low energy ions are expelled by the positive beam potential to leave behind low energy electrons that act to provide some degree of space charge neutrality. The probability is very small that these low energy electrons can combine with a high energy ion to cause deionization. Beam transport may therefore be viewed as the directed travel of a stream of high energy ions in a sea of low energy electrons having random velocity distributions. With low energy, high perveance beams enhanced means for generating such low energy electrons must be provided [106–108] or electron traps and reflectors [109] must be added to the beam line to inhibit the loss of the beam neutralizing electrons. These concepts may be applied in the mass resolving magnet, as well as elsewhere along the path of ion transport. These techniques usually involve one or a combination of the following methods and subsystems:

- Strategically placed gas bleeds that facilitate formation of ion-electron pairs near critical points along the beam path.
- System optics that maximize beam cross-section area except in critical places where a small beam cross-section is mandated.
- Subsystems that generate gas discharges near critical points along the beam path.
- Cusp-field magnetic mirror devices that are positioned to prevent the escape of neutralizing electrons.

7.4.5 Post-Analysis Acceleration and Deceleration

For many applications it becomes desirable to either accelerate or decelerate the beam ions before implanting. For example, the difficulties of transporting low energy high perveance beams suggest that a higher energy beam be extracted and mass analyzed, and then decelerated prior to implantation. As another example, CMOS well implants frequently require very high energy ions which may most

effectively be produced by acceleration to final implant energy after mass analysis. Three methods of post-analysis acceleration are employed in current generation ion implanters:

- DC acceleration/deceleration,
- Acceleration using Linear Accelerators (LINAC's),
- Acceleration using Tandem accelerators.

It is further noted that post-analysis acceleration may be preceded by beam scanning. These cases will be discussed in the section on beam scanning.

7.4.5.1 DC Acceleration and Deceleration

Figure 7.37 depicts the typical method of using an added DC potential to either accelerate or decelerate the mass resolved beam. By this method the entire source and extraction optics (with associated source and extraction power supplies) as well as the mass resolving magnet are mounted in a terminal that is maintained at a potential that is different from ground potential. After mass analysis, the beam passes out of the terminal through an acceleration gap activated by the terminal potential. Post-analysis acceleration requires that the terminal be maintained at a positive potential with respect to ground and deceleration requires a negative terminal potential. Engineering and cost considerations limit terminal potentials to ≤ 125 kV. The limits of the extraction potential and post-acceleration potentials limit the energy for singly charged ions to approximately 250 keV. With triply charged ions, energies of approximately 750 keV are achievable, usually with limited beam currents. Virtually all commercially available mid-current ion implanters from Sumitomo Eaton Nova Ltd. (SEN), Axcelis Technologies Inc. (Axcelis), Varian Equipment Associates, Inc. (VSEA), Nissin Ion Equipment Co. (Nissin) and Applied Materials Inc. (Applied) use DC acceleration. The placement of the acceleration stage with respect to the beam scanning stages differs with different system architectures and will be covered more fully in the section on beam scanning.

7.4.5.2 Radio Frequency Linear Accelerators

When energies approximately greater than 750 keV are required, other means of post-analysis acceleration are used. The NV-GSD/HE high energy ion implant products from Axcelis and SEN make use of a linear accelerator to achieve ion energies as high as 3 MeV. The linear accelerator consists of

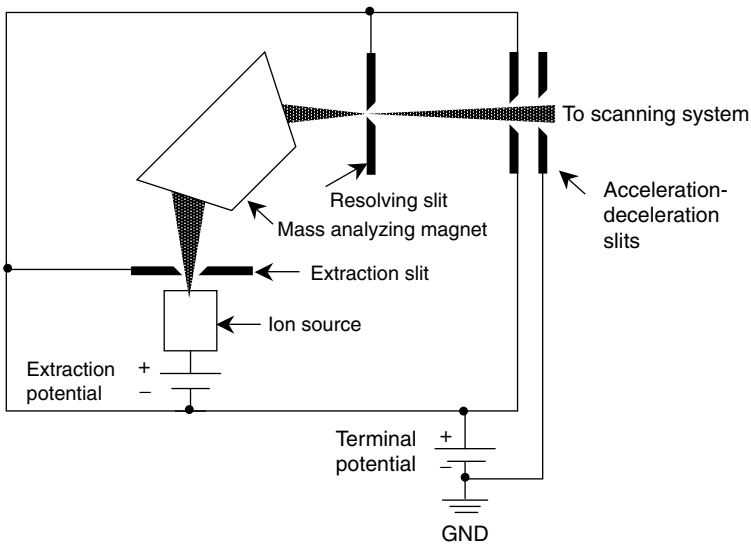


FIGURE 7.37 Schematic illustration of DC post-analysis acceleration (or deceleration).

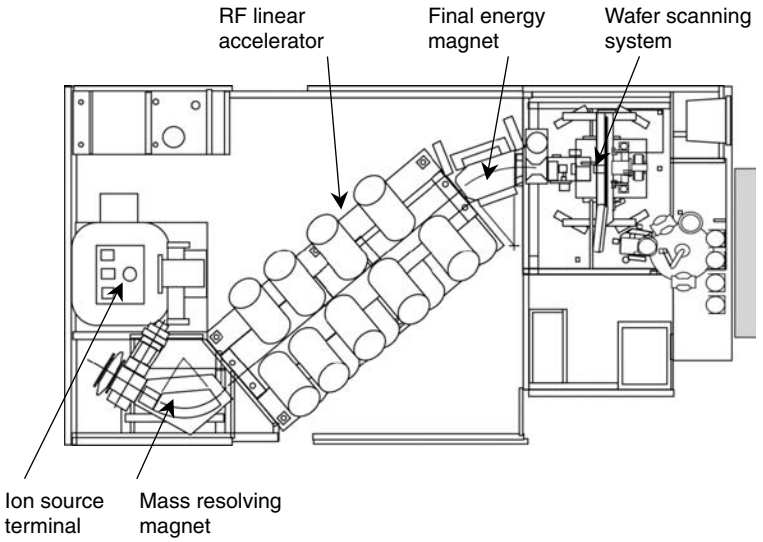


FIGURE 7.38 High energy ion implanter using linear accelerator (LINAC) ion acceleration from Axcelis and Sumitomo Eaton Nova Ltd. (SEN).

a linear series of accelerating gaps that are activated by RF voltages and are properly phased such that a bunch of ions enters into each accelerating gap at the proper phase to receive an energy boost. In this way, very high ion energies are achieved without the need for excessively large voltage drops across any accelerator gap. A plan drawing of this system is shown in Figure 7.38.

Each accelerating gap is the capacitive element of a resonant inductive–capacitive (LC) circuit. When this circuit is activated by an RF power amplifier, very large RF voltage drops ($\sim$100 kV) develop across the gap. In operation therefore, a DC beam of ions emerging from the mass analysis magnet enters the first stage of the LINAC which serves to bunch the ions. These then pass onto the first stage where the RF voltage is properly phased that the ion bunch is accelerated. This continues with each successive accelerating stage until final energy is achieved. The LINAC also requires quadrupole focusing lenses since passage through the accelerating gap tends to cause the ion bunch to diverge. The final stage of the LINAC may optionally be driven in a phase mode that results in debunching. Thereafter the ion beam passes through a final mass resolving magnet that assures that ions of the proper final energy are allowed to pass on to the scanning system. References 110 and 111 provide a more complete description of the LINAC and the simulation tool used to tune the accelerator. Reference 112 provides a more comprehensive analysis of LINAC physics.

7.4.5.3 Tandem Accelerators

Another means of post-analysis acceleration that is capable of achieving MeV beam energies makes use of a Tandem accelerator. Here the ion source and extraction system inject the ion beam into a magnesium vapor charge exchange canal, which converts positive ions to negative ions. The negative ions are subsequently mass analyzed and injected into the Tandem accelerator. The first stage of the accelerator has an acceleration column that ends in a terminal having an applied DC voltage of up to $+650$ kV. Singly charged negative ions are thereby accelerated to 650 keV and injected into an argon gas cell maintained at the terminal voltage. Here, electrons are stripped from the negative ions thereby converting these to positive ions. A second acceleration column leading from the high voltage terminal accelerates the positive ions to ground, providing another 650 keV energy boost. Final ion energies of up to 1.3 MeV are achievable with singly charged ions. The ion beam that emerges from the Tandem accelerator then

passes through a final magnetic energy filter before moving to the scanning system. References 113 and 114 provide a more comprehensive overview of Tandem-based ion implantation systems.

7.4.6 Beam Scanning System and Dose Control

The beam scanning system provides the means for uniformly painting the entire surface of the wafer with the desired ion species. Lateral control of dopant penetration into the wafer is accomplished through the use of photoresist masks, and/or the use of topographical features on the device, e.g., the use of a MOSFET gate stack to control the placement of the source and drain implants. Scanning may involve moving the beam relative to the wafer, or moving the wafer relative to the beam, or combinations of the both. As may be anticipated, all these techniques can result in variation of the incidence angle of the beam on the wafer. For nanoscale devices, such angle variations can result in undesirable device variability. Consequently, most modern implanters also provide some means of beam angle control that is incorporated into the scanning system.

7.4.6.1 Electrostatic and Electromagnetic Beam Scanning

Low perveance beams such as are present in mid-current ion implanters may be scanned by passing the beam between a pair of electrodes that are electrically activated by a scanning generator. The generator imposes a high frequency alternating potential to the electrodes which causes to- and fro-deflection of the beam. This scanning method is depicted in Figure 7.39.

The electrostatic scanning method is restricted to low perveance beams since the scanning electrodes cause neutralizing electrons to be separated from the beam. This results in unacceptable beam blow up for high perveance beams. The scanning generator typically scans the beam at approximately 1000 Hz. For such scanning systems, the wafer is also translated to cause uniform exposure of the wafer to the rastered ion flux. For example, referring again to Figure 7.39 one might imagine the wafer being translated up and down normal to the plane of figure until the entire wafer has received the required ion dose.

It is also possible to translate the wafer in a direction that makes some angle with respect to the beam thereby giving rise to "tilt" implants. These are frequently used to cause ion penetration beneath some wafer feature or mask. For example "halo" implants make use of such tilt to deliver an ion dose

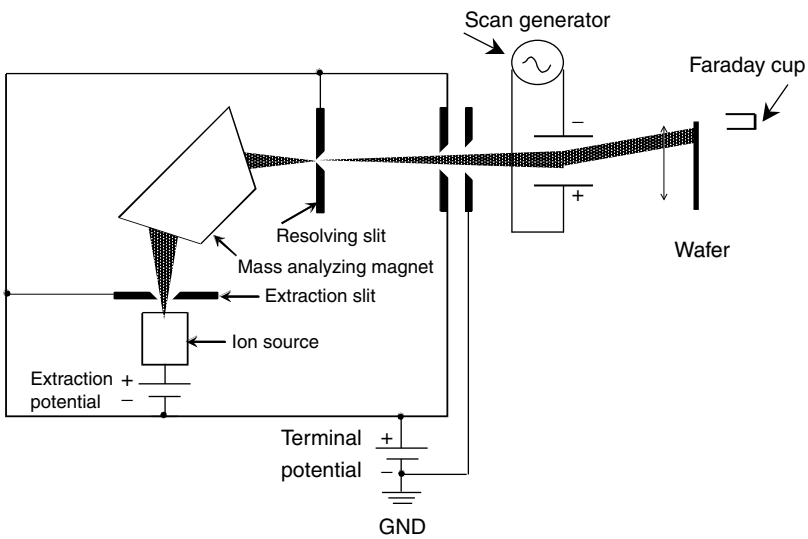


FIGURE 7.39 Electrostatic scanning of a post-accelerated ion beam.

underneath the edge of a MOSFET gate stack which serves as the implant mask. This practice assures alignment of the halo implant to the gate stack thereby avoiding a more costly photoresist masking process that is also more subject to alignment errors. This process is commonly used to reduce the sub-threshold leakage of highly scaled MOSFET devices. Implant symmetry requires that the wafer be rotated through four (or more) quadrants to assure that the implanted ions are symmetrically distributed beneath the mask. Such implants are usually called "quad" implants.

Figure 7.39 also depicts a Faraday cup positioned behind the wafer. The Faraday cup accumulates the ion beam charge for a specific (but very short) time interval, thereby providing the dose control system with regular snapshots of the beam current. The wafer translation speed may be modulated in response to the current snapshots to assure uniform exposure of the wafer to the beam. In addition, the wave form of the scan generator may be controlled to assure uniform wafer beam exposure as the beam fast-scans over the wafer.

Some attempts have been made to use alternating magnetic dipole fields to scan high perveance beams. The large inductive load of such a scanner limits the fast scan frequency to approximately 150–200 Hz, which could lead to implant dose micro-non-uniformities (striping) for low dose implants where the number of beam passes over the wafer is small. Reference 115 provides details of a high current oxygen ion implanter that makes use of such scanning on a multi-wafer batch mounted on a spinning disk. Medium Current Ion implanters available from Nissin Ion Equipment Co., Ltd, make also use magnetic beam scanning [116].

Referring to Figure 7.39 it is apparent that electrostatic beam scanning when configured as shown leads to a variation in implant angle across the fast-scan direction. Such angular variation leads to unacceptable variability in MOSFET transistor performance. As a consequence, all recently introduced electrostatically scanned ion implanters provide some means of manipulating the beam to provide parallel beam scanning. This is schematically depicted in Figure 7.40.

Both magnetic and electrostatic methods are currently in use to provide parallel scanning. The VIISta 810 medium current ion implanter from VSEA, uses a dipole magnet with rotated pole pieces together with electrostatic scanning to provide parallel scanning [117]. The Optima MD ion implanter from Axcelis uses an electrostatic lens in conjunction with electrostatic scanning to provide beam parallel scanning. Such lenses may be designed for use in acceleration or deceleration mode. Figure 7.41 depicts the Optima MD beam line with electrostatic scan and beam parallelizing lens. As is also evident from

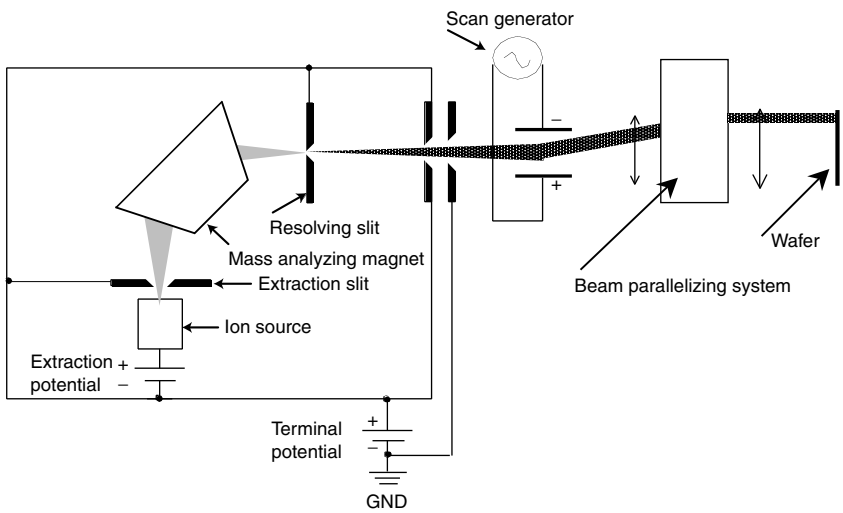


FIGURE 7.40 Electrostatic scanning with beam parallelizing optics.

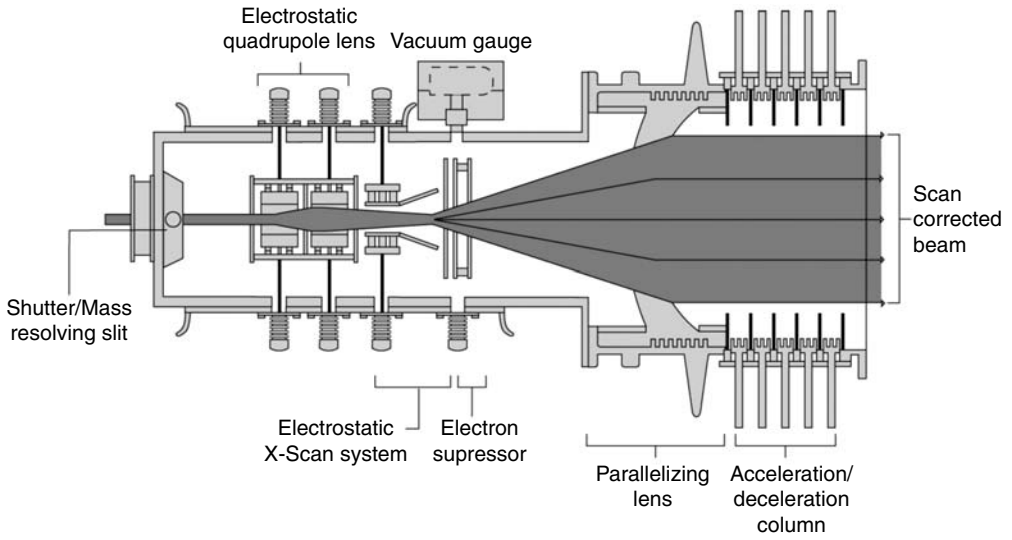


FIGURE 7.41 Optima molecular dynamics (MD) beamline with electrostatic scan and electrostatic beam parallelizing lens.

the figure, this system architecture employs post-analysis acceleration or deceleration downstream of the electrostatic scanning system.

An overview of the entire Optima MD beamline is given in Figure 7.42.

The EXCEED2300V Ion Implanter from Nissin Ion Equipment Co. uses a dipole parallelizing magnet together with magnetic beam scanning [116]. Here, post-analysis acceleration is used, after which the beam is passed through a final magnetic mass/energy filter. The beam then passes through the dipole scanning magnet and final parallelizing magnet, the latter of which is called the Collimator magnet in Ref. 116.

All of the above scanning concepts allow for ion implanter architectures that permit single-wafer implantation up to relatively high tilt angles. As mentioned earlier [103], an alternative method for single

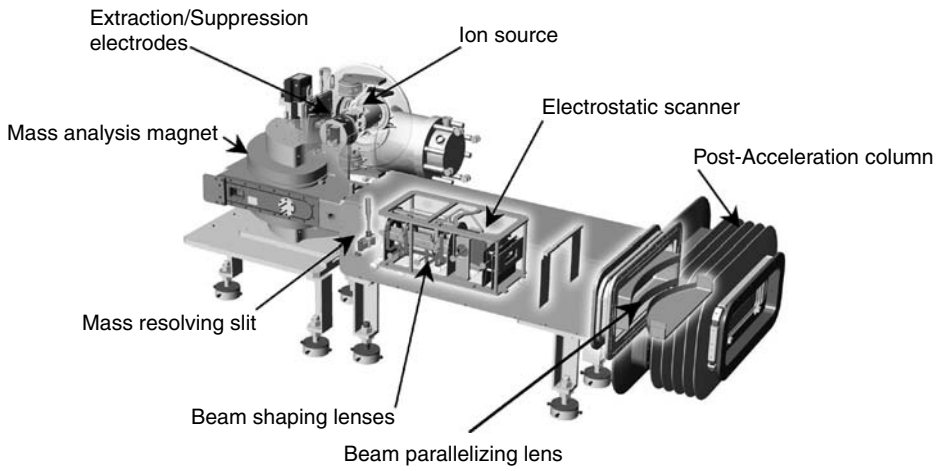


FIGURE 7.42 Optima MD beamline overview.

wafer implantation is the use of extraction optics and mass analysis systems that produce a ribbon shaped beam that is collimated using magnetic optics. Equipment based on this concept, available from Varian, has gained significant traction in the market.

The ion implantation process causes significant wafer heating, which is analyzed more fully in Section 7.5.4 of this chapter. Wafer heating is undesirable from the standpoint of implant process control, as well as photoresist mask integrity. For these reasons, all ion implanters provide some means for heat transfer from the wafer to a cooled wafer pedestal. Current generation single wafer implanters use electrostatic chucks to hold the wafer against the pedestal. Wafer-platen heat transfer is accomplished by means of a pliable silicone elastomer layer or by introducing a flow of a heat transfer gas between pedestal and wafer [118,119]. A comprehensive analysis of the physics of wafer cooling is provided in Ref. 120.

7.4.6.2 Dual Mechanical Scanning

Another way of achieving the implantation of a uniform dose is to translate the wafer in two orthogonal directions with respect to the beam. One scan direction is usually rapid to avoid dose micro-non-uniformities (striping) and wafer over-heating whereas the other is slow and frequently linked to the dose control system. Historically, high current ion implanters have used a configuration such as shown in Figure 7.43 to achieve these ends.

Here, a multi-wafer batch is mounted near the periphery of a disk that spins at approximately 1200 RPM during ion implantation (fast scan). This creates an annular shaped implantation region that has an annulus width equal to the beam height. The spinning disk is also translated in a direction that is generally orthogonal to the spin axis (slow scan) thereby causing the beam to uniformly implant a larger annulus that encompasses the entire wafer batch. A Faraday detector mounted behind the plane of the spinning disk periodically receives a beam current sample which is then used to modulate the translation speed of the disk. Dose uniformity considerations require that the slow scan distance be sufficiently great to assure that the beam entirely clears the wafer batch at each end of the slow scan travel. The wafers usually rest on wafer pads that have a coating of a silicone elastomer heat transfer agent. The wafer pad is inclined with respect to the disk spin axis. The choice of inclination angle varies with different products but varies in the range of 1.5°–7°. The centrifugal force that results from this configuration causes the wafers to be pressed against the elastomer heat transfer agent thereby assuring heat transfer from the wafer to the cooled pad. The elastomer interface also provides sufficient adhesion to assure that the

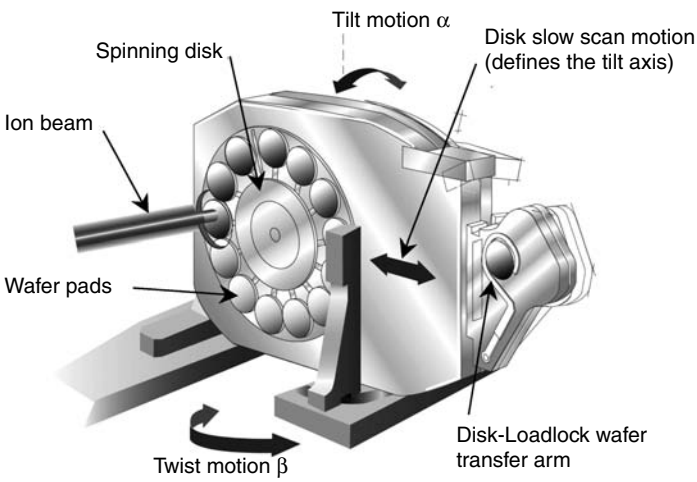


FIGURE 7.43 Spinning disk dual mechanical wafer batch scanning.

wafers remain in place during and ion implantation. Wafer restraints at the edge of the disk are required during disk spin-up and spin-down.

Batch implantation systems of this type also require that some means be provided for control of the incidence angle with which the ion beam impinges on the wafers. Usually two angles are required to control incidence angle and the beam orientation with respect to primary crystal axes of the wafer. As with single wafer implants, for many applications such as drain extension implants and halo implants, transistor device performance symmetry requires that the implant dose be divided between four equivalent angular positions. The multi-wafer implanters available from Axcelis and SEN achieve this by pivoting the disk about two orthogonal axes as shown in Figure 7.43. One axis is parallel to the slow scan directions and inclinations about this axis are called "Tilt" and the angle designated as α. Inclination about the orthogonal axis is called "Twist" and is designated as β. High current implanters available from Applied and VSEA use a different approach to achieving tilt and twist. Here the spinning disk is tilted, and the wafers rotated about an axis defined by the wafer normal (twist) prior to mounting on the disk.

As discussed in Section 7.5.6 of this chapter, wafers mounted on a spinning disk, may be viewed as translating along a circular path described the radial position of the wafer, and also rotating about an axis that is parallel to the spin axis of the disk. This latter rotation can give rise to angular variations across the wafer unless the disk spin axis is parallel to the axis defined by the ion beam. These variations limit the useable tilt angles to approximately 10° for most spinning disk systems.

It is apparent that dual-mechanical scanning allows for short beamlines. This arises because the alternate beam scanning methods require either electrostatic or magnetic manipulation of the beam which inevitably adds to beamline path length. This represents a major productivity challenge for the transport of high perveance beams. For this reason, implants that require high beam current at low energy are usually accomplished in systems that employ dual mechanical scanning. Figure 7.44 gives an overall view of a spinning disk multi-wafer implanter available from Axcelis and SEN. References 121 and 122 provide similar overviews of products available from respectively Applied Materials, Inc. and VSEA Inc.

Figure 7.44 also shows a Bernas source with extraction optics, a mass resolving magnet with beam guide designed to prevent the escape of beam neutralizing electrons, beam deceleration electrodes and the wafer disk. Implicit in this picture is a beam deceleration system as shown in Figure 7.37. The disk is labeled a "virtual slot disk" which will be explained in the Section 7.3 on dose control. The component

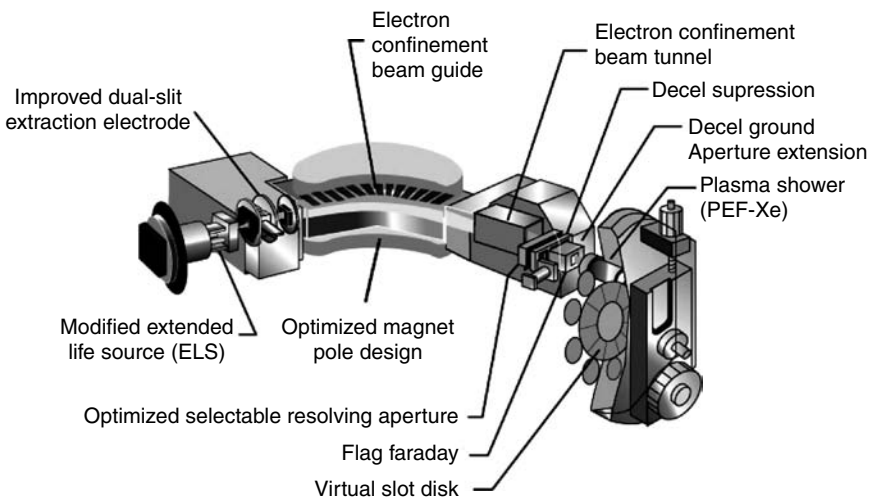


FIGURE 7.44 Spinning disk multi-wafer ion implanter from Axcelis and SEN.

labeled Plasma Shower will be discussed in Section 7.4.7 where control of wafer charging will be discussed.

7.4.6.3 In Situ Implant Dose Control

In situ dose controls are all based on the use of a Faraday detector which measures accumulated charge. Most implanter beamlines have a Faraday detector near the wafer pedestal (or multi-wafer disk), as well as upstream of the wafers. The latter Faraday detector is used as a set up monitor to qualify implant recipe changes. The component labeled "Flag Faraday" in Figure 7.44 is one example of a set up monitor. The Faraday detector used for dose monitoring is usually positioned in the plane of the wafer, or slightly behind. This detector is configured to measure charge accumulation during a specific, short, time interval, thereby determining the instantaneous beam current. This current sample is then forwarded to the dose control system that adjusts the implant process to assure a constant dose rate on the wafer. A commonly used way of accomplishing this task is by modulating the slow scan speed, e.g., a higher than expected beam current prompts an increased slow scan speed and a lower than expected beam current prompts a decreased slow-scan speed.

It is important to note that the Faraday detector measures accumulated charge, and any energetic neutral species that enters the detector will not be recorded but will contribute to the wafer dose. Neutralization of ions in the beam is commonly encountered in ion implanters and compensation for the effect of such neutrals is usually programmed into most dose control systems. This issue is covered in greater detail in Section 7.5.1 of this chapter.

A dose control system for a multi-wafer spinning disk system is depicted in Figure 7.45.

Depicted in this figure is a solid disk having wafer pads and a physical slot that is precisely machined into the disk. The solid disk blocks the ions from entering the Faraday detector except when the slot allows a beam sample to propagate into the Disk Faraday. Since the disk rotates at 1210 RPM, the Disk Faraday collects 1210 charge samples per minute. Since the slot width provides a precise definition of time, a measured beam current is determined and this current value is transmitted to the dose controller that compares the actual dose with the dose setting. The controller then modulates the slow scan speed to assure that the dose on the wafer batch remains at the dose setting.

Solid disks have been identified as a source of metal and implant species cross contamination. Consequently present generation implanters have eliminated the solid disk to replace this with a wheel

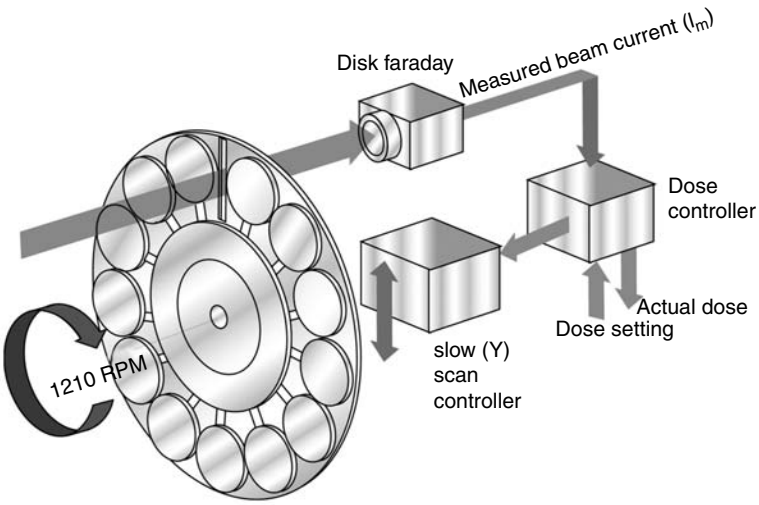


FIGURE 7.45 Spinning disk dose control system.

with wafer pedestals at the ends of spoke-shaped arms. However, this configuration still presents a known "slot" that chops the current flow to the Faraday. The known variability of the gap between wafers allows the dose control system to controller to compensate for the variable charge collection times with the result that beam current may still be accurately measured. The use of this hardware and technique has been named "Virtual Slot Disk" which has been identified in Figure 7.44. The dose control capabilities of the Virtual Slot Disk have been reviewed in Ref. 123.

Other dose control systems for single wafer and multi-wafer implanters are all based on the general practice of collecting charge for a measured time interval, deriving a beam current from the measurements and then using a dose controller to modulate the wafer transport system to maintain the required dose rate on the wafer.

7.4.7 Wafer Charging Control

The transport of an ion beam through the beamline has been likened to a directed beam of fast ions traveling through a "sea" of thermal electrons having random velocity distributions. When such a beam impinges on a wafer, a positive surface charge develops despite the fact the number density of thermal electrons may equal the number density of fast ions in the beam. This positive surface potential can give rise to a current flow through sensitive transistor structures on the electrically grounded wafer such as the gate oxide of a MOSFET. This current flow may cause outright device failure or shortened lifetime. In addition the positive surface potential can result in the extraction of neutralizing electrons from the beam resulting in beam blow up and beam strike on apertures near the wafer. Here, dose non-uniformity as well as metal surface contamination of the wafer can ensue. For these reasons, ion implant systems provide the means for introducing a compensating electron current to the wafer surface. Early ion implanters used secondary electron flood systems to provide a flood of electrons in the vicinity of the wafer surface near the beam strike area. This approach had two short-comings: firstly, the negative space charge of the electron population outside the beam inhibited the incorporation of the electrons into the positive space charge regions interior the beam, and secondly, these secondary electron floods tended to produce relatively high negative potentials at the periphery of the beam strike area thereby giving rise another device failure mode [124,125]. For these reasons, all present generation ion implanters use plasma flood systems for wafer charging control. These devices provide the means for generating a low energy plasma surrounding the beam in the immediate vicinity of the wafer. Here, in addition to providing a copious supply of low energy electrons, a large concentration of low energy plasma ions is provided. These ions help to screen the negative space charge potentials near the periphery of the beam thereby enabling better retention of compensating electrons in the positive space charge regions of the beam, and by flooding the wafer with low energy ions help to mitigate the formation of negative potentials on the wafer near the edges of the beam strike area. Reference 126 provides a quantitative overview of the beam plasma environment at the wafer surface. Further discussions of charge damage and the use of charge monitor wafers are provided in Section 7.5.3 of this chapter.

Figure 7.46 depicts a typical plasma flood system. Shown here is a conductive extension tube through which the ion beam passes. An arc chamber is affixed to the conductive tube containing a resistively heated filament cathode that is biased negatively with respect to the anodic arc chamber. Electrons emitted from the filament create a dense plasma within the arc chamber. This plasma drifts out of the arc chamber through an orifice in the arc chamber wall to create a diffuse plasma within the extension tube. The extension tube and arc chamber are further biased negatively with respect to ground by a bias power supply. The plasma potential of the diffuse plasma within the extension tube is influenced by this bias potential. Therefore, by varying the bias, the current flowing to the nearby wafer can be varied and controlled. Typically a bias in the range of 1–5 V is used. The extension tube wall also acts as an ion-electron recombination site, which may limit the achievable plasma density surrounding the ion beam. Some plasma flood systems use permanent magnets around the extension tube to create magnetic cusp-fields [127,128]. These fields inhibit electrons from diffusing to the extension tube wall thereby allowing for a higher density and more uniform plasma.

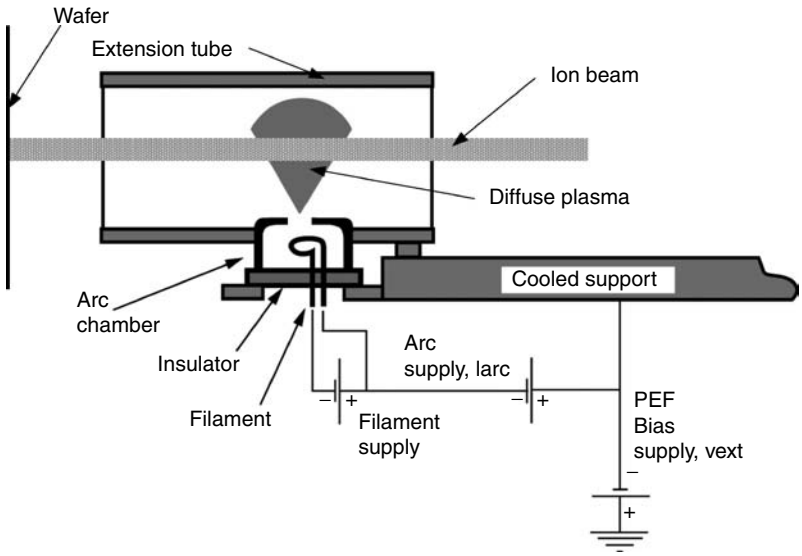


FIGURE 7.46 Typical plasma flood system for wafer charging control.

7.4.8 End Station

The end station of an ion implantation system serves to present a continuous stream of wafers to the ion beam. In order to achieve this end, a sequence of steps depicted in Figure 7.47 must be executed.

It is noted that in all modern ion implanters this sequence of steps in not serially executed. Serial execution would require prohibitively long time intervals with the consequence that wafer throughputs would be unacceptably low. All end stations have an intrinsic limit in the speed with which fresh, un-implanted wafers can be presented to the wafer pedestal. The time interval between the placement of successive fresh wafers to the wafer pedestal, t_m is one determinant of the wafer throughput limit of the end station. Another determinant is the amount of time, called overhead time t_{OH} that is required after wafer pedestal placement for the wafer to be in position to receive the ion beam. The final determinant is the implant time t_{imp} that is required to uniformly implant the wafer to the required dose. If the implant time is sufficiently short then the inequality of Equation 7.43 applies and the ion implanter is said to be operating at the mechanical limit.

$$t_{OH} + t_{IMP} \leq t_m \tag{7.43}$$

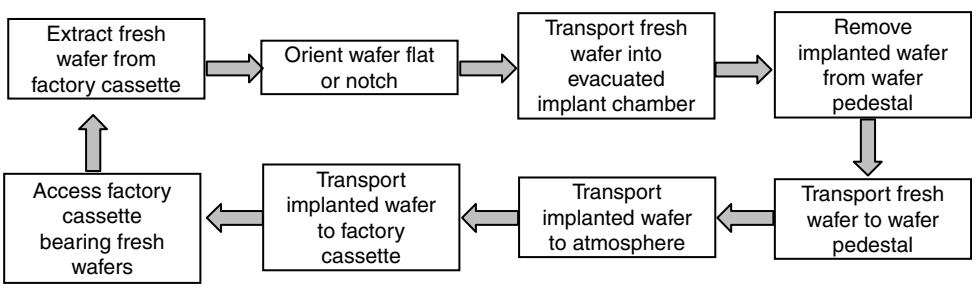

Extract fresh wafer from factory cassette
Orient wafer flat or notch
Transport fresh wafer into evacuated implant chamber
Remove implanted wafer from wafer pedestal
Transport fresh wafer to wafer pedestal
Transport implanted wafer to atmosphere
Transport implanted wafer to factory cassette
Access factory cassette bearing fresh wafers

FIGURE 7.47 Ion implanter end station wafer handling execution steps.

Modern ion implanter all have mechanical limit throughputs that are greater than 230 wafers/h. If the implant time is longer, then the inequality of Equation 7.44 applies and the ion implanter is said to be operating in the dose rate limited regime

$$t_{OH} + t_{IMP} > t_m \tag{7.44}$$

The implant time is determined by the beam current, the ion charge state, the required dose, and the implant area. This is given by Equation 7.45.

$$t_{IMP} = \frac{ADZe}{I} \tag{7.45}$$

Here A is the implant area in cm^2, D is the dose in atoms/cm^2, Z is the ion charge state, e is the electronic charge in Coulombs, and I is the beam current in Amperes. For singly charged ions $Z=1$, doubly charged $Z=2$, etc. The resultant implant time is given in seconds. The implant area is not the wafer area but is the total area that must be implanted in order for the wafer to receive a uniform dose. The beam utilization efficiency is just the ratio of the wafer area to the implant area, and is equal to one in the ideal case where the implant area is equal to the wafer area. For most modern ion implanters the beam utilization efficiency is approximately 60%.

The above considerations give rise to wafer throughput vs. dose characteristics that are depicted in Figure 7.48. Two cases are depicted here. Case A is for an implanter that has a lower mechanical limit and achieves this with a smaller implant overhead time. Case B is for an implanter that has a higher mechanical limit but requires a greater implant overhead time. It is noted that that the greater overhead time of Case B, causes transition to the dose limited regime at a lower implant dose. The productive advantages/disadvantages of these two different cases is dependent on the dose at the transition point, and the overall implant recipe mix that the implanter is expected to execute. The message here is that an implanter with a higher mechanical throughput is not necessarily the most productive implanter. Overall productivity estimates must take into account the overall recipe mix, and the other equipment characteristics such as beam current capability, and beam utilization efficiency.

Another important characteristic of the implant system end station is the vacuum pumping efficiency of the implant chamber and nearby beamline. Implantation with photoresist masks results in the copious emission of gases, comprised primarily of hydrogen [129]. This emission is cyclical because different

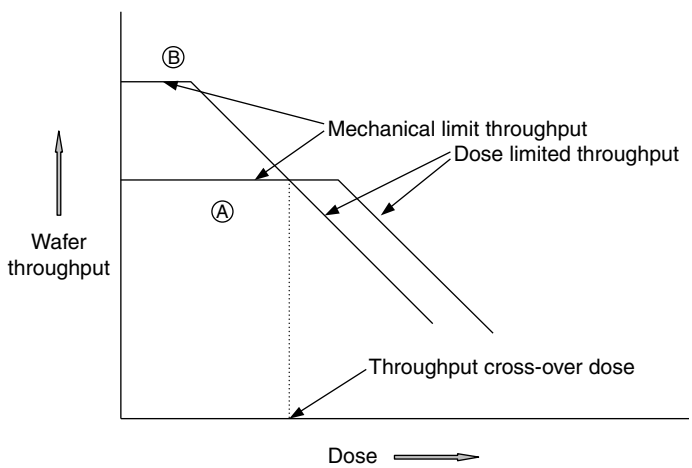


FIGURE 7.48 Typical implanter throughput curves for two different mechanical limits and implant overhead times.

degrees of photoresist exposure to the beam occur according to whether the beam scans across a wafer diameter, or across a chord of the circular wafer. In addition, the gas emission decays as the resist becomes carbonized. Pumping efficiency comes into play because the gas emission, by raising implant chamber pressure increases the likelihood of ion–neutral interactions that can result in a change in the charge state of the implanting ion. At low energies, a fraction of the positive ions can become converted to fast neutral atoms with virtually the same kinetic energy. These are not detected by the Faraday detector and thereby cause a dose error. In addition, dose shifts across the wafer may arise. For example, if the system uses a dipole magnet to create parallel scan, then an ion that is neutralized in the parallelizing magnet will implant at an some off-angle that is determine by the point in the dipole magnet where neutralization occurred. Beam parallelizing systems that use electrostatic beam parallelization are also vulnerable beam neutralization errors. The result is that cross-wafer dose error and implant angle errors ensue unless a corrective system is put in place or very efficient vacuum pumping is provided. Dose errors of this type are discussed more fully in Section 7.5 of this chapter. In most commercial ion implanters this problem addressed by one or more of the following equipment strategies:

- The employment of very efficient vacuum pumping systems that minimize pressure rise.
- Provision of conductance limiting apertures where the beam enters the implant chamber which limit pressure rise in the beamline upstream of the implant chamber.
- The use of pressure-dose compensation algorithms.
- Provision of a final filter that removes neutrals and other off-energy species from the ion beam.
- The use of algorithms that limit beam current, or otherwise cut-off the ion beam if the implant chamber pressure rises above a predetermined limit.

7.5 Process Control in Ion Implantation

7.5.1 Introduction

Ion implantation is a complex process. It involves the generation of energetic beams of ionized atoms or molecules, which are accelerated to a predetermined energy, and made to impinge on target materials having a variety of crystal structure and orientation with widely different surface properties, structures, and coatings. This complicated combination of equipment and process variables often yields implant process results that differ from expectations. Thus, to repeatedly achieve desired process results, it is important to understand the physics and materials interactions that underlie the key implant process interactions. This knowledge may then be used to implement appropriate process controls and procedures supported by the relevant metrology, failure cause analysis, and corrective action.

The list of the most important implant process parameters that must be controlled and monitored by the implant process engineer is extensive. It includes: control of implantation dose, depth, angle, and uniformity; control of contaminants introduced by the implant process; control of ion channeling; managing charge buildup on insulating structures such as gate oxides; and, finally, dealing with various manifestations of ion beam heating and interaction with photoresist masks. Process interactions are numerous and complex, and the reader is encouraged to refer to proceedings of the biannual conferences on ion implantation technology available through the IEEE. In this chapter we will focus on: ion beam residual gas interactions and the resultant dose errors and control, ion beam wafer charging damage detection and control, ion beam photoresist interactions and the impact on resist mask stripping, implant angle control and contamination control.

7.5.2 Beam-Residual Gas Interactions Control of Implanted Dose

A common cause of implant dose errors is the charge exchange between ions in the beam, and atoms and molecules of residual gas in the implanter beamline [130]. This phenomenon is especially important in implanters where the total beam power into a photoresist coated wafer results in

significant pressure rise in the implant chamber. Here, high current and high energy implanters are the most susceptible to dose error. Mid-current implanters have traditionally not been as susceptible but with the drive for increased implanter productivity, many mid-current machines now provide beam currents where pressure induced dose error are becoming manifest. Charge exchange can result in the neutralization of the ionized particle (for low and medium energy ions) or an increase in the ion charge state due to electron stripping (for high energy ions). Since virtually all implanters use accumulated ion charge measured by a Faraday detector to measure accumulated dose, changes in charge state will result in dose errors, e.g., neutralized ions will not be detected and over-dosing will occur. The momentum and energy of the accelerated dopant atom remains virtually unchanged from such interactions which implies that charge exchange interactions that occur upstream of the mass analysis magnet will be removed. However, interactions that occur downstream of the mass analysis magnet can result in dose as well as energy errors. For those implanters that do not use post-analysis ion acceleration, the deviations are confined to dose errors. For those implanters that use post-analysis acceleration or decelaration, both energy and dose errors may occur. It is also noted that systems which use magnetic or electrostatic beam parallelizing components (see Section 7.4.6.1 of this chapter) are subject to dose shifts across the wafer as well as the implantation of off-angle components. A common problem with ion implanters that use post-analysis deceleration to produce very shallow p–n junctions, is the contribution from more energetic fast neutrals that yields an unwanted deep tail to the implant profile.

At low to medium ion energies, neutralization of beam ions is the dominant charge exchange mechanism. This can be described by the 1D differential equation:

$$dn = -n\sigma N dx \tag{7.46}$$

where n ion flux incident on the differential charge exchange volume having length dx, σ is the cross-section of the charge exchange reaction, N is the volume concentration (molecules/unit volume) of residual gas atoms in the charge exchange region, and x is the length of the beam path into the charge exchange region. The initial point for charge exchange is usually taken to be the mass resolving slit of the beamline. Calling this point $x=0$, and the flux at this point n_0, Equation 7.46 may be integrated to yield the expression

$$n_L = n_0 \exp[-\sigma NL] \tag{7.47}$$

where n_L is the charged particle flux measured at the Faraday detector, and L is the total path length from the mass resolving slit to the Faraday detector. Since n_L gives the measured dose and n_0 gives the true dose, it is helpful to rewrite Equation 7.47 to get an expression that is explicit in n_0.

$$n_0 = n_L \exp[\sigma NL] \tag{7.48}$$

In Equation 7.48, the initial beam current I_0 and measured beam current I_L may be substituted for n_0 and n_L, respectively to give:

$$I_0 = I_L \exp[\sigma NL] \tag{7.49}$$

Equation 7.49 is not in convenient form for dosimetry correction since the path length L, from mass resolving slit to Faraday detector is a constant, and N, a function of the pressure, varies with implant conditions. For example, when implanting through a photoresist mask the pressure varies with the scan position of the beam, rising to a maximum with a wafer diameter scan, and falling to a minimum for an edge scan. For this reason, a more convenient form of Equation 7.49 is achieved when N is expressed as a function of the beamline pressure P. Using the ideal gas laws where $P=NkT$, where k is Boltzmann's constant, and T is the absolute temperature, permits the substitution of N, giving the equation:

$$I_0 = I_L \exp[KP] \tag{7.50}$$

where K is defined by the expression:

$$K = \frac{\sigma L}{kT} \tag{7.51}$$

Notice that since the beamline gas temperature is essentially a constant, and since L is fixed by beamline design, K is a constant and may conveniently be used to compensate for charge exchange using Equation 7.50.

The residual gas in the beamline downstream of the mass resolving slit arises mainly from two sources: photoresist outgassing and gas from the plasma flood system. Plasma flood systems are used to control wafer charging and typically make use of argon or xenon. For these reasons, the beamline pressure has a constant component from the flood gun, and a cyclical component that arises from scanning of the photoresist coated wafer. Beamline pressure near the wafer varies from the baseline value, typically in the upper range from 10^{-7} Torr to the 10^{-4} Torr when implanting resist-coated wafer. If the beamline pressure is continually monitored, and the K value predetermined, then the measured pressure and Equation 7.50 can be used for real time correction of the charge collected by the Faraday detector to yield the true dose. Modern high current implanters have incorporated automated look-up tables to determine the value of K in Equation 7.50, eliminating the need to perform a measurement for every implant condition. When pressure in the beamline is near the baseline value, such as when implanting wafers without resist and flood gun, then ion beam neutralization is negligible and the current measured by the Faraday detector accurately represents the actual dopant flux from the ion beam.

At higher energies, electron stripping becomes an important source of dose error. The precise energy range where this becomes important varies with the ion specie but for ions with relatively small atomic number such as Boron and Phosphorus, it becomes significant in the range from 0.5 to 1 MeV. The various theories of electron stripping and the energy range where it becomes important are discussed in Ref. 130. The simplest case is one where some fraction of the incident ions in charge state q_1 is raised to a charge state q_2. Here, two equations similar to Equation 7.50 describe the stripping process and both charge states contribute to the measured Faraday current. This simple case yields a beam current correction as given in Equation 7.52 below.

$$I_L = I_0[1 + (\gamma - 1)(1 - \exp(-\sigma NL))] = I_0[1 + (\gamma - 1)(1 - \exp(-KP))] \tag{7.52}$$

where γ is the ratio of the initial to the final charge state, e.g., the stripping of one electron to go from $a+1$ to $a+2$ charge state gives a value of $\gamma = 2$. The modeling of more complex scenarios [130] gives equations of the same form with the exception that the term $(\gamma - 1)$ becomes a more complicated function of the reaction cross-sections, and different K values are introduced into the exponential term. However, Equation 7.52 is suitable for dose error compensation with appropriate choices of the parameters γ and K. This approach has yielded excellent dose correction over a wide range of pressures. Optimum parameter values are tabulated in the look-up table accessible by the computer controlling the dosimetry system, and are invisible to the user. It is important to note that in the case of beam neutralization, when the final charge state is zero ($\gamma = 0$, e.g., charge neutralization), Equation 7.53 becomes equal to Equation 7.50. Dose correction based on Equation 7.52 is employed in the GSD-HE and GSD-VHE high energy implanter available from Axcelis.

High current Implanters from Applied Materials and Varian rely on different approaches for dealing with charge exchange reactions between the beam and the residual gas. The multi-wafer high current implanter families from Applied Materials employ an open loop dose control system. Here, the beam current is measured at the end of the implant disk slow scan cycle, when the beam is off the wafer, and when photoresist outgassing effects are minimal. When the beam impinges on the wafer batch and photoresist outgassing leads to pressure rise, the dose rate is assumed to be remain constant, and the

beam current readings from the Faraday detector are used only to monitor beam dropouts (glitching) and other critical errors.

In the VIISion family of multi-wafer high current ion implanters from Varian, the end station with the process disk and charge neutralization (plasma electron flood) hardware are the Faraday charge detector. Thus, charge exchange between the ions and residual gas that occurs in the end station has minimum impact on implanter dosimetry, since the total charge in end station (neutralizedion plus ionized molecules of the residual gas) remains unchanged, and is registered by the implanter dosimetry system. Dosimetry systems utilizing this approach allow continuous dose monitoring, but are electrically and mechanically more complex.

With the increased use of single wafer implanters linked to the industry transition to 300 mm wafers, beam-residual gas charge exchange adds complicating factors. New generation dual mechanical scanned single wafer implanters are in the early product introduction stages. Here, dose errors similar to those encountered in dual mechanical scanned multi-wafer are encountered and can be compensated as previously discussed. In addition, a variety of systems are available that use either ribbon beam or scanned beam approaches. These are discussed in greater detail in Section 7.4 of this chapter. For systems that use beam scanning, a beam parallelizing system must be provided to assure that the beam incidence angle on the wafer remains constant as the beam is rastered across the wafer. Both magnetic and electrostatic methods are used both of which are sensitive to the charge state of the ion. For such systems, beam/residual gas charge exchange can lead to dose error, and if such exchange takes place in the beam parallelizing system dose shifts across the wafer become a problem. In addition, off-angle fast neutrals can impact on the wafer causing undesirable degradation of transistor performance. The favored approach in dealing with these issues is to limit end station pressure rise. This can be achieved by limiting beam current and associated end station pressure rise, or by interrupting the implant process if end station pressure rises to a pre-determined limit. These approaches can implant implanter productivity. The alternative is to increase end station pumping speed, which significantly adds to system cost. The Optima™ MD implanter from Axcelis uses an electrostatic energy filter downstream of the beam-parallelizing lens. This eliminates ions that have been neutralized in the beam parallelizing system, thereby eliminating off-angle and off-energy species. However pressure induced charge exchange that occurs downstream of the energy filter can still lead to dose shifts.

There is an additional opportunity for dose error and energy contamination that is encountered when implanting doubly-charged ions. In general, the problem can be encountered for any species that can form molecular dimers, e.g., P_2 and As_2, as with phosphorus and arsenic. Here the ion source produces singly and multiply charged ions (e.g., P^+, P^{++}, As^+, and As^{++}) as well as ionized dimers (P_2^+ and As_2^+). In this environment it is possible to generate ionized species having different energies but the same magnetic rigidity. Consequently these will pass through the mass resolving magnet and create dose error and energy contamination. From Section 7.4.3.2 of this chapter the magnetic rigidity of an ion species is given by:

$$RB = \sqrt{\frac{2mE}{q^2}} \tag{7.53}$$

In the above equation m is the ion mass, E its energy, and q its charge, in the mass resolving magnet. Consider now the extraction and implantation of doubly charged phosphorus as an example. Here, along with the extraction of the P^{++} ion, the singly charged dimer P_2^+ is also extracted. The energy of the doubly charged phosphorus ion is $2qV$, and that of the dimer is $1qV$, where V is the extraction potential. The dimer ion may also be disassociated through collisions with residual gas molecules in the extraction region. This results in the following disassociation reaction:

$$P_2^+ \xrightarrow{\text{collision}} P^+ + P^o$$

Since the energy of the original dimer particle is equally partitioned between the two equal mass fractions, the energy of the resultant P^+ ion is $qV/2$, i.e., it has one fourth the energy of the P^{++} ion. As a consequence, two charged particles of equal mass are injected into the mass resolving magnet: a singly charged ion having energy $qV/2$, and a doubly charged ion having energy $2qV$. From Equation 7.53 it can be determined that these two particles have the same magnetic rigidity, and will therefore be equally transmitted through the mass resolving magnet. The implantation depth of these two ion species differs because of their different energy. Since one species is a singly charged ion, and the other doubly charged, dosimetry errors also ensue. The use of multiply charged ion implantation is widely used in mid-current and high-energy ion implanters. High-energy implanters that use LINAC such as the Axcelis Paradigm and Paradigm-XE machines avoid this problem since when the LINAC is tuned to transmit and accelerate the doubly charged species, it will not transmit the singly charged contaminant. Mid-current systems use additional filtration to eliminate the singly charged contaminant [131].

Dimer implantation has become increasingly popular because it increases implanter productivity for the manufacture of very shallow drain extension, and source–drain junctions. Implanting a dimer gives rise to the possibility of energy contamination as well. This energy contamination arises as follows: when tuning the singly charged dimer beam, doubly charged atomic ions extracted from the source may capture one electron prior to entering the analyzer magnet. Possessing the same magnetic rigidity as the dimer, these singly charged atomic ions have $4\times$ velocity per ion compared to singly charged dimer atoms, and are referred to as "4-x energy contamination." The effect may be minimized through choice of proper source operating conditions. Detailed studies of the electrical performance of dimer implants have been reported in the literature for low energy implants [132,133]. These studies demonstrated equivalent performance for all device properties in comparative studies with monomer implants.

7.5.3 Charging and Gate Oxide Integrity

As discussed in Section 7.4 of this chapter, implantation of wafers with energetic positive ion beams, may cause the electrostatic potential of the wafer surface to readily reach tens or even hundreds of volts, unless the surface and the beam is properly compensated with electrons. Electrostatic potentials of less than ten volts on the surface of a charge-sensitive structure such as the MOSFET gate stack or a DRAM storage capacitor can induce charge carrier tunneling through the dielectric layer that causes either dielectric breakdown, or *wearout*. Both failure modes result from the buildup of latent defects in the dielectric layer generated by the current flow. Low levels of these defects result in the reduction of dielectric life expectancy under normal device operating conditions. High levels can cause immediate device failure.

Device scaling mandates the continuous reduction of MOSFET gate oxide thickness with the consequence that the charge induced current flow and thereby the sensitivity to dielectric wearout increases. Similarly scaling of the DRAM storage cell area requires the use of thinner dielectric layers. In both cases scaling has required the replacement of the pure silicon dioxide dielectric layer to SiO_2 having profiled nitrogen additions distributed throughout the film thickness. Future scaling requires the replacement of these layers with other materials having higher dielectric constants [134] that have different susceptibility to charge damage.

Electron tunneling induced failure has been the subject of investigation in terms of both dielectric failure and error detection [135]. As discussed in Ref. 135, two different classes of tunneling are known. For thicker films Fowler–Nordheim tunneling is the relevant tunneling process whereas for thinner films direct tunneling applies. The breakdown electric field for SiO_2 is approximately 12–13 MV/cm and can be reached at less than 3 V beam potentials for 1.2 nm thick gate dielectric layers typical of 90 nm CMOS devices. This imposes stringent requirements for beam space charge neutralization, and requires a reduction of electrostatic beam potential to maximum of several volts. To provide this degree of control, most modern ion implanters have charge sensors integrated into the implant chamber to measure the positive and negative potential profiles produced by the scanned beam. A description of such sensors and their operation is provided in Ref. 136.

A number of factors in addition to beam potential profile and dielectric layer thickness, influence gate dielectric sensitivity to charging. Among them are the I–V characteristics of beam plasma [137], and photoresist pattern on the wafer [138], the charge sensitivity of the device, the type of doping (*n*- or *p*-) under the gate oxide, the ratio of the gate charge collection area to the gate area, and gate geometry (exposed sidewall vs. sidewall spacer). These complexities make it virtually impossible to accurately predict the correlation between device yield and measured beam characteristics. Therefore to determine the upper limits for positive and negative swings of beam potential that will ensure no device yield loss it is necessary to establish an actual correlation between device yield and charge monitor response. In addition it is necessary to control the beamline parameters that determine the beam potential profile.

As discussed more fully in Section 7.4 of this chapter, the beam potential profile is influenced by the residual gases that are introduced into the beamline to produce electrons by a process first described by Holmes [139]. In addition many ion implanters have plasma flood systems that provide an added current of low energy electrons and ions to minimize the positive and negative beam potential excursions. These plasma flood systems also require a flow of a plasma forming gas such as argon or xenon. Day to day monitoring of the beamline and plasma flood background gas pressures along with charge monitor responses are valuable process control tools that can be used to establish process baselines and control limits which can assist in control of beam potential profiles and to trigger further investigation if limits are exceeded. However, actual charging damage can only be obtained by either measuring the yield of actual devices, or by using charge-sensitive short loop test structures. Given cost and turnaround time considerations, charging damage monitoring using test structures is the preferred and most commonly used approach.

Traditionally short loop test structures for the measurement of gate dielectric integrity (traditionally called GOI for gate oxide integrity) have used parameters such as charge-to-breakdown (Q_{BD}), voltage-to-breakdown (V_{BD}), and time-to-breakdown (T_{BD}). These measurement techniques are all based on the known capacity of high quality gate dielectric layers to typically pass a cumulative ≈ 10 Coulombs/cm^2 of charge prior to breakdown [140]. Charge damage introduced to the gate during implantation will contribute to this total allowed charge and will therefore result in a reduction of the charge that can be transported through the dielectric layer during subsequent measurements. Thus, by measuring a difference between Q_{BD} on the unimplanted and implanted structures, it is possible to evaluate charging damage introduced during ion implantation. In addition, as referenced above, very thin oxides allow direct electron tunneling and are not subject to the same failure mechanisms as thicker oxides where Fowler–Nordheim tunneling applies.

In the typical MOS capacitor structure used for Q_{BD} or similar measurements, a charge collecting antenna usually made from polysilicon forms the first capacitor plate attached to the thin gate dielectric layer. As later will be described more fully, the substrate or the control gate of an Electrically erasable programmable read only memory (EEPROM) device forms the second plate. The use of large area antennas increases the charge collection area of the test structure and improves its sensitivity to the charging effects. Several other slightly different designs of test capacitors are also commonly used [141]. They differ by having either perimeter-intensive or area-intensive antennas, that are protected or exposed to the beam at the edge of the gate stack. Usually, up to several hundred clusters of MOS capacitors with different gate antenna area ratios are fabricated on a single wafer. This allows for the matching of the test structure array sensitivity to the large variety of implantation conditions that require evaluation. In this manner, test wafers with a variety of charge-sensitive structure can be used for the characterization of wafer charging at various beam currents, doses, and energies. The requirement for having a large number of similar capacitors on the test wafers is dictated by the statistical nature of the charging damage. Typically, half of the test capacitors on a wafer are measured prior to implantation and the remaining after the implantation. The statistical analysis of collected data is required for accurate consistent interpretation. The data collected from test structures are usually presented as a Weibull plot (*Y*-axis has an ln(-ln(1-*F*) scale)), (*F* is the cumulative failure rate), and *X*-axis is the parameter tracked (Q_{BD}, V_{BD}, or T_{BD}). As shown in Figure 7.49 the Weibull plot for charge-to-breakdown is a format that readily allows for comparisons between different test conditions.

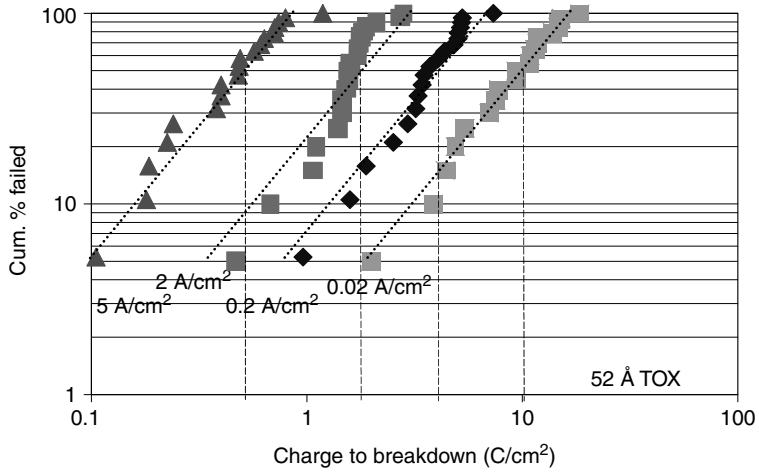


FIGURE 7.49 Typical Q_{BD} Weibull plot of the cumulative failure rate of metal oxide semiconductor (MOS) under various stress conditions.

SPIDER and CHARM®-2 [142] wafers are two commercially available test devices frequently used for characterization of wafer charging. As shown in Figure 7.50, the CHARM®-2 structure is an E^2PROM device with a large charge collection antenna attached to the control gate [143]. Post-implant threshold voltage shifts in device permit determination of the peak potentials induced by the ion beam. CHARM®-2 wafers contain clusters of charge-sensitive devices distributed on the wafer. Individual structures in each cluster have charge collection antennas connected to the substrate, with shunts having a different resistivity. This allows the determination not only of peak beam potentials, but also of I–V characteristics of the beam plasma. This latter characteristic provides essential information about the electric current passed through gate dielectric, the ultimate cause of charging damage. An analysis of peak voltages recorded by the E^2PROM devices located in different areas on the wafers allows the mapping of beam potential across the wafer.

SPIDER structures consist of CMOS devices fabricated to the first metal level, with large charge collection antennas connected to each gate stack [136,144]. Implant charging damage is estimated from the deviations in the electrical characteristics displayed by the individual PMOS and NMOS field effect transistors. One of the advantages of using SPIDER structures is that charging damage is gauged by

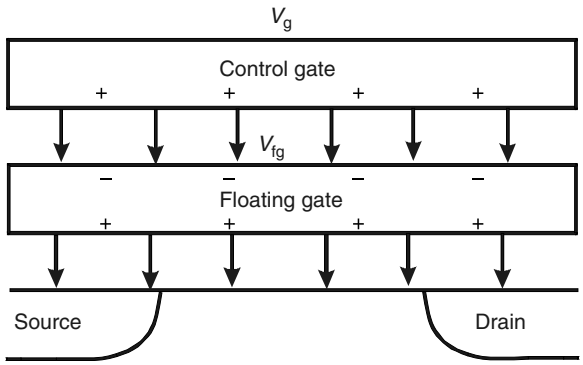


FIGURE 7.50 CHARM®-2 charge flux sensor.

monitoring of the very same electrical characteristics that real devices would sustain when exposed to the implantation process. Post-implant threshold voltage shifts are usually measured. Here, beam-induced charge carrier tunneling through the gate dielectric generates defects in the bulk of the dielectric and at the dielectric/channel interface. A certain fraction of these defects can trap electrons and holes, resulting in a threshold voltage shift. The degree of charging damage can be determined since higher charging stress leads to increased defect generation and to a greater threshold voltage shift. Unfortunately, some defects remain neutral and therefore do not contribute to a V_t shift. Information about these electrically inactive defects is also important because these may trap electric charge during device operation, or otherwise manifest themselves in reduced device life expectancy. The concentration of neutral and active dielectric/channel interface defects can be measured on the SPIDER structure by means of a charge pumping technique [145,146]. Since these measurements are non-destructive, a variety of techniques can be applied to the same test device permitting for a more comprehensive understanding of the problem.

7.5.4 Wafer Cooling and Photoresist Mask Integrity

The high beam powers achievable in high current and high energy implanters can substantially increase the wafer temperature unless effective cooling is provided. Increased wafer temperature can compromise the implant process in a number of ways. For example, as reference to Section 7.2 will show, it causes a reduction in the rate of implantation damage accumulation. For crystalline materials, this increases the dose required for amorphization [147], thus compromising the quality and repeatability of pre-amorphizing implants commonly used to facilitate the formation of shallow, highly activated dopant junctions in silicon. This thermal induced variability can in turn cause run-to-run deviations of the implanted dopant distribution. Elevated temperature may also result in the formation of undesirable defect complexes with high thermal stability and in extreme cases, may trigger radiation-enhanced diffusion of the implanted dopant [148]. These phenomena can affect both depth distribution and activation of implanted atoms, thus compromising implant process control.

A more common problem results from the fact that even moderate wafer heating can cause problems with the photoresist masking layers. When the beam power raises the photoresist temperature to the vicinity of its glass transition temperature T_g, critical dimensional (CD) changes will begin to occur. At temperatures at or above the T_g the resist will flow, causing failure of the masking layer. Other resist heating effects include: burning [149], blistering, flaking, and popping [149,150], the latter two generating high particle levels and potential machine contamination. Figure 7.51 shows an example of photoresist blistering (a), popping (b) and a redeposited photoresist flake from a resist-popping event (c). At implant doses over about 1×10^{15} ions/cm^2 the photoresist surface layers exposed to the ion bombardment loose significant H and O and become transformed into a dense carbonized layer that is compressively stressed and caps the unexposed photoresist [151–153]. If, as this dense layer is formed,

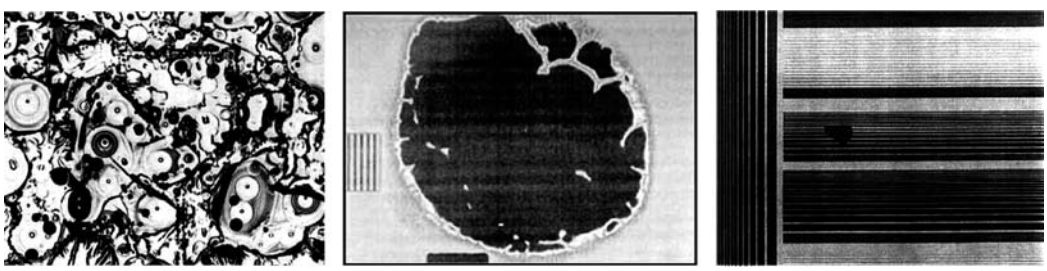


FIGURE 7.51 Implanted resist problems, (a) resist blistering, (b) resist popping, (c) a redeposited resist flake after popping.

the underlying resist experiences a temperature increase to near its T_g (as in the implantation process for example) the resist will yield to the compressive stress in the carbonized layer, causing a mini explosion on the wafer [149]. An explosion or resist popping event occurs when this highly stressed, carbonaceous crust, erupts violently as gas pressure builds underneath this crust layer and the layer tries to relieve this compressive stress. One solution to this popping problem is to UV photostabilize the photoresist causing the T_g to dramatically increase [154,155]. Figure 7.52 compares a UV photostabilized (a) vs. a hard-baked (b) photoresist implanted and heated to approximately 130°C. The hardbaked resist explodes violently while the photostabilized wafer shows no signs of popping or blistering. UV photostabilization can raise the T_g of resists to about 250°C [154].

To prevent photoresist popping in the implanter it is important to keep the peak photoresist temperature well below T_g throughout the duration of the implant. The maximum implanted wafer temperature rise depends on a variety of factors, including: (1) total ion beam energy and current, (2) ion beam current density, (3) ion dose, (4) efficiency of wafer cooling, (5) beam dwell time or scanning speed, and (6) the degree of thermal conductivity between the resist surface and the wafer. For most low and medium current implants with a beam power (beam voltage times beam current) of less than 30 W, resist heating is generally not a significant issue. However modern high current and high-energy implanters can deposit a kilowatt or more of beam power (500 W is typically a maximum useable beam power where resists can be used as implant masking materials). Therefore high current and high-energy implanters, require the provision of advanced wafer cooling strategies.

To understand wafer heating effects one must realize that wafer heating occurs on at least three different time scales of: (1) seconds, (2) milliseconds, and (3) microseconds. The peak temperature the photoresist masking layer reaches is a sum of these three heating regimes. We will examine each of these separately. "Average" wafer heating occurs on the "seconds" time scale. This heating is a balance between the beam power supplied to the wafer and the ability of the wafer chuck to cool the wafer. Well-designed wafer cooling systems can dissipate between 15 and 30 mW/cm^2 and typically employ backside He gas cooling to improve cooling efficiency. For high implant doses the average wafer temperature reaches a mean steady state value equal to the temperature where the input power approximately equals the power dissipated through the chuck or:

$$T_R = \frac{P_i}{P_0 A} \tag{7.54}$$

where T_R, temperature rise of the wafer (°C); P_i, input beam power (watts); P_0, power dissipated by chuck (watts/cm^2/°C); A, wafer surface area (cm^2).

For a 50 keV, 10 mA beam (beam power = 500 W), chuck temperature of 25°C and a chuck cooling efficiency of 30 mW/cm^2/°C the average wafer heating reaches 48.6°C. Figure 7.53 represents a finite

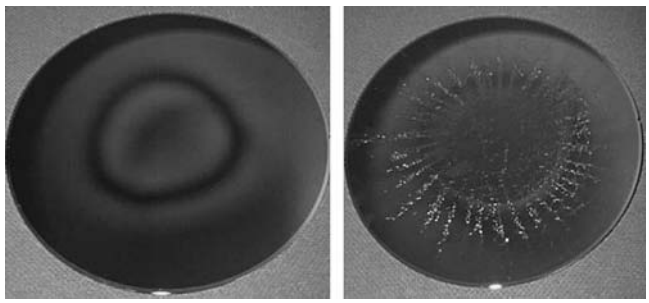


FIGURE 7.52 Ion implanted (80 keV As, 1×10^{16} ions/cm^2) wafer heated to 130°C after implantation. (a) UV photostabilized to 230°C, (b) hardbaked to 110°C.

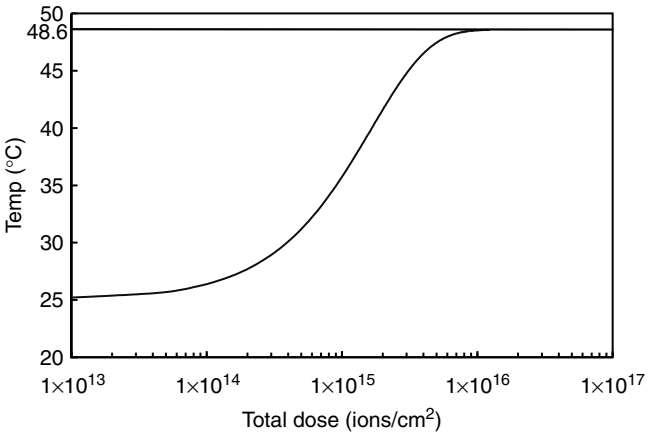


FIGURE 7.53 Simulation of the average temperature rise in a single wafer implanter for a 300 mm silicon wafer with 50 keV, 10 mA implant conditions and a chuck contact resistance of 30 mw/cm^2 °C.

difference computer simulation of the mean wafer temperature as a function of implant dose for a 50 keV, 10 mA beam impinging on a single wafer implanter. At doses greater than approximately 1×10^{16} atoms/cm^2 the mean wafer temperature reaches the value predicted by Equation 7.54.

Transient wafer heating occurs on the millisecond timescale. The portion of the wafer directly under the beam experiences localized heating. For most acceptable implant conditions (those for which the local heating is below about 100°C) the heat loss is dominated by lateral heat conduction in the wafer. This condition was described by Pitaway and is given in Equation 7.55 [156].

$$T_r = \frac{Q\sqrt{kt}}{K\sqrt{\pi}}\left(1-e^{\frac{-R^2}{4kt}}\right)+\frac{QR}{2K}\text{erfc}\left(\frac{R}{2\sqrt{kt}}\right) \tag{7.55}$$

where T_r, temperature rise (°C); Q, heat input (watts/cm^2); k, thermal diffusivity of silicon (0.909 cm^2/s); K, thermal conductivity of silicon (1.49 W/cm–°C); t, beam dwell time (s); R, beam radius (cm); erfc, complementary error function.

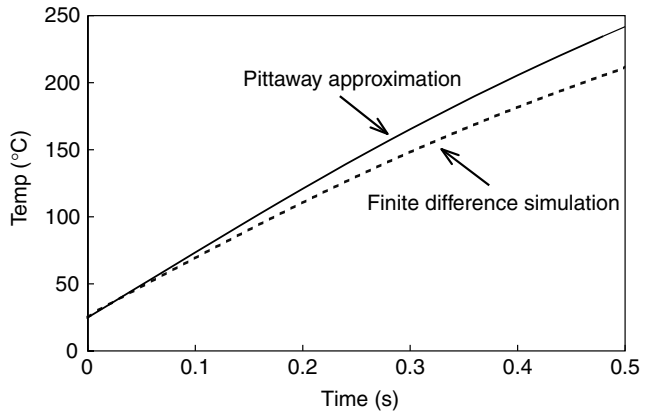


FIGURE 7.54 Local wafer heating from a 50 keV, 10 mA beam impinging on a silicon wafer vs. beam dwell time. (Assuming initial wafer temperature of 25°C and a beam radius of 3 cm).

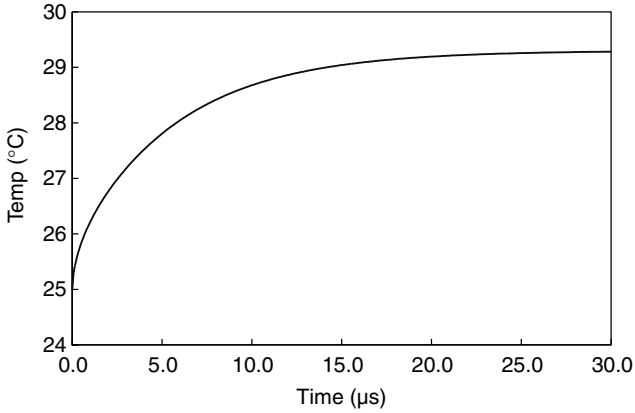


FIGURE 7.55 Finite difference thermal analysis of the temperature rise at the surface in a 1 μm thick (i-line) photoresist due to a 50 keV, 10 mA, 3 cm diameter beam vs. time. The resist develops a thermal gradient of about 4°C/μm due to the poor thermal conductivity of the resist.

In Figure 7.54, a finite difference thermal analysis of an ion beam heated wafer vs. beam dwell time is shown, for the same 50 kW, 10 mA beam, assuming a circular beam diameter of 3 cm. To maintain transient local heating below about 50°C, beam dwell needs to be below about 50 m/s. To achieve this, implanters rapidly scan the beam, wafer, or both.

Dynamic resist heating occurs on the microsecond time scale. This heating is characterized by creation of a thermal gradient in the photoresist where the resist surface heats above the local wafer temperature. This gradient can be as high as 10°C/μm for high beam powers and highly insulating photoresists. The magnitude of resist heating depends on the thermal properties of the photoresist, photoresist thickness, ion beam power, and ion range. To a smaller extent resist pattern density, number and thickness of underlying insulating layers, etc., contributes to variations in resist heating. Figure 7.55 is a finite difference analysis of an un-patterned, 1 μm thick, i-line photoresist on silicon subjected to the same 50 keV, 10 mA beam vs. time. In this analysis, a temperature gradient of about 4°C/μm is developed in the resist in just a few microseconds. These temperature gradients can cause changes to the resist sidewall profile if the mean resist temperature approaches the glass transition temperature, T_g. Such sidewall changes can cause implant deviations in the region near the altered resist sidewall.

7.5.5 Removal of Heavily Implanted Photoresists

As may be inferred from Section 7.2, energetic ions cause dramatic modifications to organic materials in addition to the aforementioned Joule heading effects. Ion/material electronic interactions, and to a lesser extent, nuclear interactions, cause bond breakage along the ion path. At low total doses ($\leq 1\times 10^{13}$ ions/cm^2) the polymeric material undergoes two competing mechanisms: chain scission and cross-linking. For many polymers (polymethylmethacrylate for example) and light ions such as boron [157], the scission process predominates over the cross-linking process and the average molecular weight of the polymer is reduced during the initial stages of the implantation process. The scission process also results in cleaving of volatile groups from the resist with the evolution of gaseous by products [158,159], such as H_2, O_2, and CO, CO_2, the majority of which is H_2 [160]. At a sufficiently high dose (greater than $\sim 1\times 10^{15}$ ions/cm^2), little if any hydrogen and oxygen remain resulting in a compressively stressed [151,153], carbonized layer, doped with the implanted ion species. The thickness of this carbonized layer has been studied [153,160,161] and is found to be about twice the projected range of the ion. Thus after implantation, three distinct regions exist in the photoresist mask: (1) a highly stressed, carbonized and

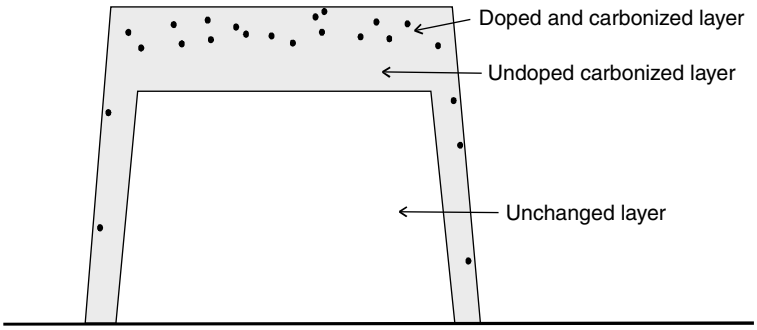


FIGURE 7.56 Schematic cross-section of an implanted photoresist feature. Three distinct layers are formed in the photoresist structure: (1) a doped and carbonized crust, (2) a secondary, undoped but carbonized layer, and (3) an unchanged layer. (After Fujimura, S., Konno, J., Hikazutani, K., and Yano, H., *Jpn. J. Appl. Phys.* 28, 2130, 1989.)

ion doped crust region, (2) a carbonized but un-doped middle region, and (3) unchanged photoresist. This is depicted in Figure 7.56. The carbonized region is a consequence of the implantation process itself, and unlike resist heating which can be reduced by implanter design and implantation conditions, this cannot be prevented.

The removal of this three layer structure presents an interesting challenge for the post-implant plasma strip process. At low to moderate stripping temperatures, the carbonized layer is highly resistant to chemical attack and the underlying encapsulated unchanged photoresist layer will volatize at the temperatures needed for reaction of the carbonized layer with the atomic oxygen stripping gas. The activation energy for atomic oxygen reaction with the carbonized layer has been measured [162] and was found to be about 2.6 eV in comparison to the activation energy for i-line photoresist of 0.17 eV (see Figure 7.57). Joshi et al. [163] have measured the oxygen activation energy of carbon materials and have found that similar activation energies are obtained for diamond-like carbon (DLC) films. This is not too surprising in that one way to form DLC films is to subject organic materials to ion bombardment [164,165] (Table 7.13).

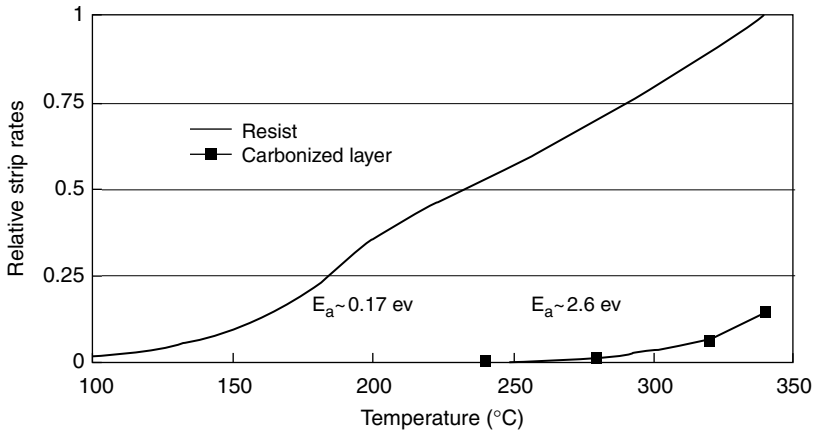


FIGURE 7.57 Relative removal rates of standard i-line photoresist and the implanted carbonized crust layer as a function of temperature for a oxygen plasma without ion bombardment. Activation energy (E_a) has been calculated from the temperature dependence of the reaction.

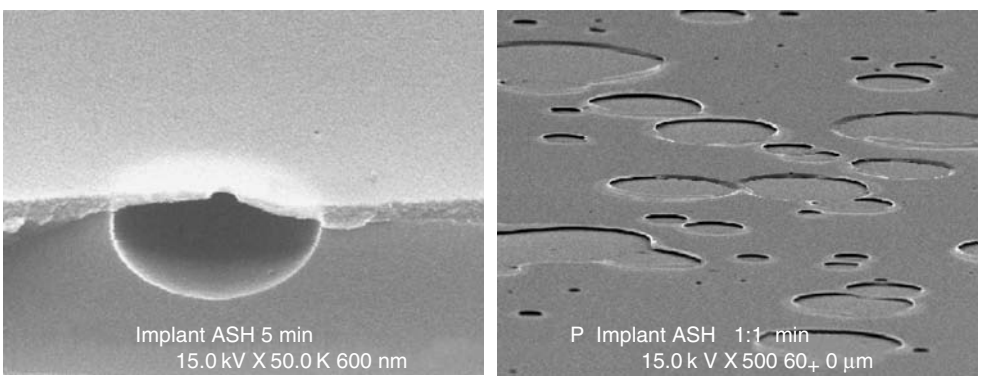
TABLE 7.13 Measured Oxygen Activation Energies for Carbon Materials

Material	Activation Energy (kJ/mol K)	Activation Energy (eV)
Carbon (graphite)	176	1.83
Carbon (pyrolytic)	164	1.70
Natural diamond	172	1.79
DLC (C2)	318	3.30
DLC (C3)	308	3.20

Source: After Joshi, A., Nimmagadda, R., and Herrington, J., *J. Vac. Sci. Technol. A*, 8, 2137, 1990.

Viewing the removal (plasma ashing) of the implanted photoresist layer on a microscopic level, one finds that the removal is very non-uniform. Figure 7.58 shows a 1.2 μm thick i-line resist, implanted with 1×10^{16} ions/cm^2 of P at 30 keV, subjected to microwave downstream oxygen plasma at 120°C. Since the reactivity of the carbonized crust layer is very small in O* at 120°C (see Figure 7.57), ashing of the photoresist begins at defects in the crust layer, allowing the active oxygen to penetrate the carbonized crust and attack the unchanged photoresist underneath. As the ashing proceeds, the crust layer looses thermal contact with the wafer and begins to heat, increasing the reactivity. Some of this overhanging crust material will unfortunately break off and fall back onto the wafer surface, regaining thermal contact, and become difficult to remove by the oxygen plasma as shown in Figure 7.59. Removal of these remaining residues is typically done in a strong acid wet clean such as an SC1 or "piranha" [166]. A common approach to reducing the residues is to ash the photoresist in a multi-temperature step process, consisting of: step (1) ashing at moderate temperatures (below the resist T_g) to generate sufficient crust breakup to prevent resist popping, then, step (2) raising the temperature to 250°C–300°C to enable sufficient reaction rate of the oxygen with the remaining carbonized material [167]. Other approaches include: (1) the use of hydrogen based plasma chemistries to breakdown the carbonized crust layer, followed by an oxygen plasma to remove the remaining photoresist [161,168], (2) using ion bombardment to enhance the reactivity of the carbonized crust [153,169,170], or (3) the use of fluorine containing gas additives to enhance the chemical reactivity of the carbonized material [168,171].

An important issue is the control of the total substrate loss that occurs during the implanted photoresist strip and clean process. Celler et al. [172] measured the silicon loss from a standard SC1 clean to be about 25–30 Å, which for ultra-shallow junctions will be problematic. Additionally the ion implantation process itself causes significant damage to the substrate that the etch rates are substantially



(a) (b)

FIGURE 7.58 Partial removal of P implanted (1×10^{16} ions/cm^2, 30 keV) resist in oxygen at 120°C after 5 min (a) and 11 min (b).

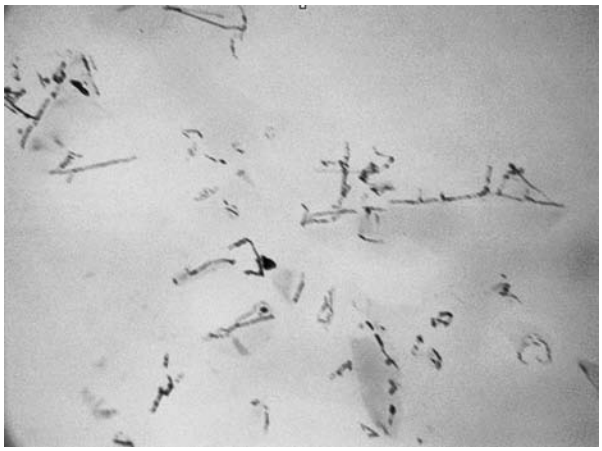


FIGURE 7.59 Residues left on wafer after oxygen ashing of heavily implanted photoresist.

increased by up to 5 times that of undamaged oxide [173]. The year 2003 ITRS [174] has flagged the substrate loss issue as a critical challenge for sub 65 nm generation device production.

7.5.6 Implant Angle Integrity

The incidence angle that the ion beam makes with the wafer is important as it relates to the crystal axes of the silicon wafer substrate, and also as it relates to shadowing produced by topographical features on the wafer such as resist masks and MOSFET gate stacks. In ion implantation, the placement of the implanted dopant depends (among other things) on the tilt (θ) and twist (ϕ) of the wafer during a given implant (Figure 7.60). Tilt is defined as the angle between the ion beam vector and the normal to the wafer surface. Twist is defined as the azimuthal angle between the projection of the ion beam on the wafer surface and the [011] crystallographic direction (the direction from the center of the wafer to the wafer notch). For $\phi = 0$, the ion beam is oriented away from the wafer normal in the direction of the wafer notch. As ϕ increases, the direction of the incoming beam relative to the normal rotates counterclockwise away from the notch, as shown in Figure 7.60. Depending on the energy of the implanted ions, errors in either θ or ϕ may result in either channeling or shadowing of the implanted ions, or both. Either condition distorts the desired 3D dopant distribution, which may adversely affect device performance. Under certain conditions, angle errors as small as 0.2° can be problematic for advanced devices.

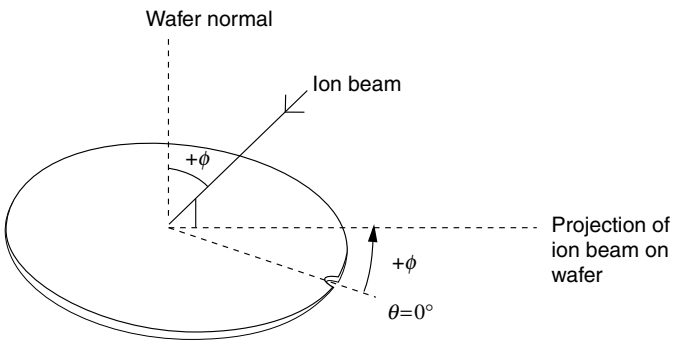


FIGURE 7.60 Definitions of tilt (θ) and twist (ϕ).

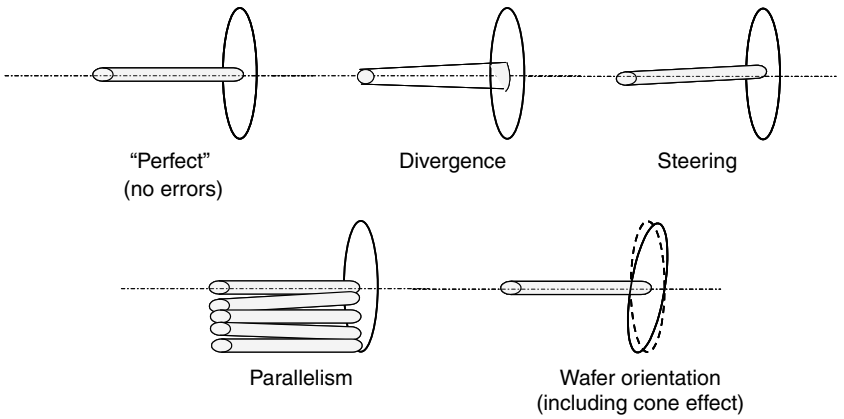


FIGURE 7.61 Pictorial representation of potential ion beam angle errors.

Shadowing effects depend on device structures only, and are not sensitive to crystal cut variations. Although not implanter-related, variations in the crystal cut of the wafer (the angle between the ⟨001⟩ direction and the wafer normal) can cause run-to-run variations in dopant profiles due to channeling. The current SEMI wafer specification dictates that the crystal cut error must be smaller than $\pm 1.0°$. In practice, most device quality wafers have crystal cut errors $< \pm 0.5°$.

The remaining sources of ion beam angle error described in this section are implanter related (Figure 7.61). The most common occurrence is an error in the beam centroid angle with respect to the wafer surface. This error can occur for stationary beams (Figure 7.61), scanned beams, or ribbon beams. It can occur in either the *x*-direction (dispersive plane for most commercial beamlines) or the *y*-direction (non-dispersive plane). See Section 7.4.4.1 for a definition of the dispersive vs. non-dispersive planes of the beamline. Steering charged particles down a beamline is roughly analogous to steering photons through a series of lenses. Every acceleration or deceleration gap acts as a lens for the ions. If any fringing electric or magnetic fields are present at the gap, there is the potential for the ions to be steered off the intended trajectory.

The ion beam can leave the source at an incorrect angle if the extraction optical elements are not carefully tuned. This angle error can propagate all the way to the implanted wafer. If the beam or any portion of the beam enters a steering element at the incorrect angle, it will leave that element with an incorrect angle, usually with a larger angle error. In this way, angle errors may be magnified as a beam progresses down a long beamline. Several techniques exist for in situ measurement of beam angle. The most common is to place a precision mask in the path of the ion beam, and determine the angle error from the location of the resulting shadow. The maximum beam steering error in either the *x*- or *y*-directions is difficult to estimate, because it depends on so many subtleties of any particular beamline. In practice, centroid steering errors of 0.5°–3.0° are commonly observed.

It is also possible for the wafer to be mechanically misaligned with respect to the frame of the implanter. As with beam centroid errors, misalignment can occur in either the *x*-direction or the *y*-direction. The magnitude of these errors is usually less than 0.5°, although errors of 1°–1.5° have occurred. Although both the beam steering and the wafer positioning offsets are referenced to the implanter frame for simplicity, what matters from a process point of view is the sum of these offsets, the beam angle to wafer normal offset. This offset can be corrected by repositioning the wafer in both the *x*- and *y*-directions, if the true combined offset angle is known and the implanter has the capability to adjust wafer position to correct beam angle offsets.

Because the beam is composed of positively charged ions, coulombic repulsion can result in beam expansion as it travels down the beamline. This may result in increasing beam divergence (Figure 7.61).

Divergence results in individual ions entering the wafers at different angles, while the average beam angle is usually not altered. Consequently, divergence effects are difficult to measure, either in situ or through device parametric tests. The spread in individual ion angles across a beam depends on (among other things) the species, energy, beam current, initial size of the beam, beamline length, and the degree of space charge neutralization in the beam. The latter factor is usually dominant especially with perveance beams; beams with the highest degree of space charge neutralization show the lowest divergence. The calculation of beam transport and associated divergence effects is a complex topic beyond the scope of this chapter. See Ref. 175 for a more complete discussion.

The issues discussed above are applicable to virtually all ion implanters, regardless of scanning architecture. Other sources of beam angle error are unique to specific implanter designs, as discussed below. Early medium current implanters employed simple electrostatic scanning of a "pencil" (fixed) beam in both the *x*- and *y*-directions. No attempt was made to compensate for the change in beam angle due to the scanning. Consequently, the angle errors were approximately:

$$\theta_E \sim \tan^{-1}[r/d] \tag{7.56}$$

where r is the wafer radius and d is the distance from the center of the scan plates to the wafers. In practice, it was common for this effect to produce angle errors of $\pm 3^\circ$ across a 150 mm wafer (the largest wafer size that these systems were equipped for). For a given location on the wafer, the magnitude of the angle error in both the *x*- and *y*- directions is roughly linear with the distance from the wafer center. Although no longer manufactured, implanters of this type are still in use in some fabs. Despite this deficiency in angle control, these implanters have been used to fabricate transistors as small as the 350 nm node.

Most modern single wafer implanters use hybrid scanning. A pencil beam is scanned in the *x*-direction, while the wafer is mechanically scanned in the *y*-direction. In between the scanning mechanism and the wafer are one or more ion lenses to bend the beam so that ion trajectories in the beam are parallel, regardless of position. The parallelizing lenses prevent the gross angle errors inherent in the early scanning mechanisms. Smaller angle errors still exist (Figure 7.61) as it is impossible to steer all the beamlets on exactly parallel trajectories.

Beam parallelism errors can arise from many sources. Every steering element in the beamline is a lens, and has the potential to alter the beam trajectory from the intended path. Compared to a pencil beam, a parallel scan beamline has more steering elements, and thus a larger potential for angle errors. Other possible sources of errors in this implanter type include:

- Errors in the precision of the power supplies in the extraction and beam steering lenses
- Mechanical errors associated with the installation of the beam steering lenses
- Any horizontal or vertical errors in the beam centroid angle prior to entering the scanning mechanism.

Depending on the implanter and the quality of the beam setup, the magnitude of these errors ranges from ± 0.1 to $\pm 2.0^\circ$ or more. Unlike a beam centroid offset, parallelism errors cannot be corrected by adjusting the wafer position.

The last source of angle errors we consider is a geometric effect particular to multi-wafer implanters employing a pencil beam and 2D mechanical scanning. The scanning is achieved by rotating a disk holding the wafers in front of the ion beam, while simultaneously scanning the disk in a linear fashion. If the axis of the disk rotation (which is typically varied to achieve the desired tilt angle) is not parallel to the ion beam vector (which is fixed), the rotation of the disk will cause variations in tilt and/or twist across the wafer as it scans through the beam with rotation of the process disk (Figure 7.62). Because most multi-wafer implanters have wafer pedestals that are tilted up from the disk surface (for wafer cooling reasons), this is sometimes referred to as the "cone angle effect."

The angle errors from this effect are extremely consistent from run-to-run and do not depend on the beam setup, because this effect is purely geometric. Additionally, tilt variations can be calculated from a

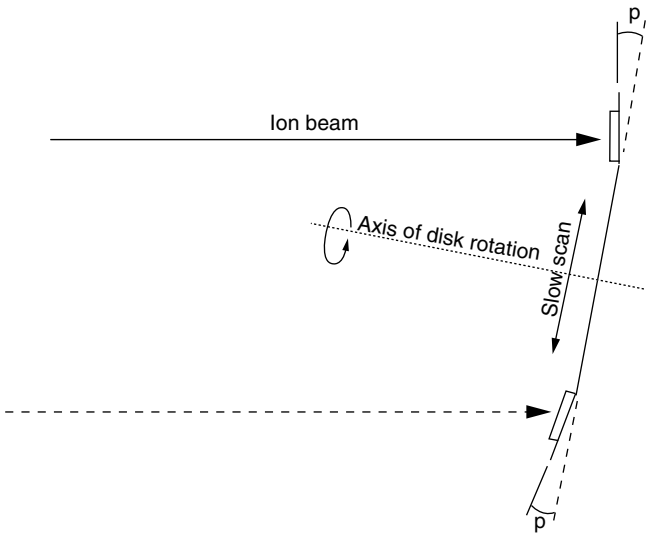


FIGURE 7.62 Illustration of changes in ion beam incident angle as the disk rotates. The wafer is normal to the ion beam at the top of the figure, but not to the imaginary parallel beam vector at the bottom.

basic understanding of a particular end station design [176,177]. While the exact calculation is complex, geometric angle errors are generally reduced by maximizing the number of wafers on the disk, minimizing the angle between the wafer pedestals and the disk plane, and minimizing the angle between the axis of rotation and the ion beam vector. Tilt errors from this effect are about ± 0.1 to $\pm 1.2°$ or more, but can exceed $\pm 3.0°$ on some older implanters

7.5.7 Wafer Contamination

The bombardment of solids with energetic ions causes sputtering, which may be responsible for elemental and particle contamination of the wafer during ion implantation. Because of this, the materials that are used in beamline construction are carefully chosen to minimize any contaminants that may be introduced into the beam. The most common sources of these contaminants include:

- Sputtered beamline components
- Portions of the wafer pedestal and support structure not covered by wafers
- The Faraday charge detector
- Residual implant species in the process chamber, deposited along with photoresist on chamber walls.

The most common contaminants include aluminum, carbon (from beamguide graphite liners or photoresist) and residual dopants from prior implant steps, usually boron, phosphorus, or arsenic. Typical surface concentration of these contaminants in modern high current batch implanters ranges from less than a hundred to up to ten thousands of parts per million (ppm) of the implant dose, which is equivalent to 10^{11}–10^{13} cm^{-2} for a 10^{15} cm^{-2} implant dose. The lower range is typical for implanters with a silicon-coated process disk [178] and where effective disk cleaning procedures are used to suppress cross-contamination. Control of these contaminating species include using low sputter yield liners and coatings, introduction of beam dumps for capturing off-axis analyzed species, and elimination of line-of-sight pathways on which sputtered material can accumulate. Additionally, control of pumping and thermal cycling during processing can reduce deposition on surfaces that result in particles and wafer

contamination. Transition metals comprise another group of contaminants. These originate from beam line materials and are found on the surface of implanted wafers in much lower quantities, typically less than 5 ppm. These species are all but eliminated through the proper use of liners and shields to prevent exposure of any metals to the beam are related beam-plasmas, such as seen in the extraction source and plasma flood systems. The impact of various contaminants on device performance and process stability strongly depends on the kind of contaminant and whether the contaminant resides on the wafer surface, or penetrates into the volume (in the case of energetic species) of the semiconductor [179]. Defining tolerable limits for each class of contaminant depends on the specifics of the process flow, the presence or absence of intrinsic gettering structures, and the type of device structures produced.

Sputtered atoms arrive at the wafer with low energy and typically reside on the surface of implanted wafers. The use of screen oxides that are grown on the wafer prior to ion implantation and removed before the post-implant anneal effectively remove these contaminants from the wafer [160]. Energetic contamination usually occurs when the dopant beam is contaminated with ions having a magnetic rigidity close to that of the dopant ions. The most common manifestation of this problem is contamination of $^{49}BF_2^+$ dopant beam with $^{98}Mo^{++}$ ions [180,181] that emanate from the molybdenum arc chamber of the ion source. Here, sputtering and ionization of the arc chamber wall results in a finite concentration of Mo^{++} ions (usually up to 10–20 ppm) in the source plasma. The use of tungsten arc chambers eliminates this contamination source. Contamination of $^{31}P^+$ ion beam with $^{31}BHF^+$ ions is another example of possible energetic contamination [182] which is mitigated by proper species change protocols. Finally, energetic contamination with carbon, oxygen, nitrogen and hydrogen can be observed during implantation into photoresist-coated wafers in ion implanters having a post-analysis acceleration stage. This happens when gases released by photoresist outgassing are not effectively evacuated from beamline, and thereby allowed migrate to the post-acceleration region, where ionization there through interaction with the ion beam ensues. These contaminants are subsequently energized by the post-acceleration potential and caused to impinge on the wafer. The projected range of these energetic contaminants is comparable to the projected range of the dopant ions, and for this reason screen oxides cannot effectively suppress this contamination. Careful optimization of implant conditions, and subsequent process monitoring are required to minimize energetic contamination and keep it below the tolerable limit. The migration of high current implants into the low energy regime has significantly reduced the amount of outgassing observed from photo-resist coated wafers, due to the shallow penetration of the implanted species into the photoresist. This, along with improvements in vacuum system design, effectively minimizes the contamination mechanism.

It is common practice in many fabs to implant multiple dopant species using the same ion implanter. Under these conditions, dopant species cross-contamination can potentially arise. The most significant contributor to cross-contamination is sputtering of the dopant atoms from the process disk and the Faraday detector that have been previously contaminated by prior implanter operation [183]. Since these are surface contaminants they may be effectively removed from the wafer surface by the use of screen oxides, or by rigorous cleaning of the wafer prior to thermal activation. In most practical cases the effect of cross-contamination shows up as a shift in sheet resistance (R_s) of the activated wafer. In situations when phosphorus is the cross-contaminant, in addition to the R_s shift, a significant increase of junction depth can occur [182]. This is due to higher phosphorus diffusivity, as compared to other common dopants. Several schemes have been devised to deal with this contamination, including the introduction of selectable strike plates within the disk Faraday. These plates are exposed to only one dopant type, thereby eliminating "cross-species" contamination of the wafer that emanates from the Faraday during implant.

Gold, iron, copper, nickel, chromium, and other transition metals are the least tolerable contaminants. All these elements introduce mid-gap traps and can therefore act as charge recombination centers. Several transition metals including gold, copper, and iron also show very high diffusivity in silicon. If these contaminate the wafer surface, they diffuse during subsequent thermal processing steps to a depth much greater than the dopant, and thereby create charge traps (deep levels) in depletion regions of the p–n junctions. When present above critical concentration, deep traps significantly increase junction leakage.

Tolerable levels of transition metal contamination depend on requirements associated with specific devices (CMOS logic, DRAM, BiCMOS, etc.), device generation, and manufacturing process flow [184]. Although under lab test conditions most devices do not demonstrate degradation of electrical characteristics if transition metals are in a range of 10^{11}–10^{12} cm^{-2}, 5 ppm is a currently accepted upper limit for these impurities in implanters. Upcoming device generations push this limit down even further, or will require more effective gettering methods.

7.5.8 Metrology

The large variety of parameters that characterize the implanted layers dictates the need for multiple analytical techniques, to support initial setup and sustain ongoing control of the implant process. These techniques provide information on chemical, electrical and structural properties of the implanted layers. Some of these techniques such as ThermaWave (TW) and four-point probe (FPP), are fast and inexpensive to use and have therefore found widespread use for routine monitoring of the implant process. For initial set up or in cases where process deviations are observed and a more comprehensive analysis is warranted, techniques such as Total x-ray Fluorescence (TXRF) and Secondary Ion Mass Spectroscopy (SIMS) may be invoked. These are destructive, more complex and time consuming.

ThermaWave [185,186] and FPP [187] are the most common techniques for implant monitoring. Both are fast, relatively inexpensive and enable wafer mapping. Thus, they are ideal tools for determining dose non-uniformity across the wafer. TW measures changes in light reflection from the wafer resulting from the implantation-induced crystal damage. TW is well suited for monitoring implants with doses starting from $\sim 10^{11}$ cm^{-2} up to amorphizing doses ($\geqq 5 \times 10^{14}$ cm^{-2} depending on ion species). Implant damage depends on dose, dose rate, implant wafer temperature, and beam/wafer orientation (channeling effects). It is therefore necessary to maintain run-to-run consistency of all implant parameters in order to draw reliable information from TW maps.

Four-point probe measurements require post-implant anneal in order to activate implanted dopant. For these reasons, particular care must be exercised over the thermal activation process to ensure that this does not contribute a measurement error.

Total x-ray Fluorescence analysis [188] is an established technique for quantitatively determining elemental surface contaminants. This technique makes use of an x-ray beam that strikes the wafer surface at a grazing angle. Interference between the incident and reflected x-ray beams compresses the incoming x-ray beam within a thin surface layer. This limits the excitation depth to several atomic layers thereby suppressing background flourescence and giving very high sensitivity to surface contaminants. The TXRF technique is best suited for surface analysis of elements having high atomic weight. The TXRF has a 10^{10} cm^{-2} or better detection limit for titanium and higher atomic weight elements. The detection limit for potassium and calcium is 10^{12} cm^{-2} and rapidly deteriorates further for lighter elements.

The SIMS technique [189] is a destructive method for depth-dependent elemental analysis. It is time consuming, requires a skillful operator and is more expensive to use than TW and FPP techniques. On the other hand, the SIMS technique has low elemental detection limits (sub-ppm level or 10^{12}–10^{16} cm^{-3}), a wide dynamic range of impurity concentration measurements (4–5 orders of magnitude), depth resolution as low as 2 nm and high lateral resolution (down to sub-1 μm). The SIMS measurements allow the detection of virtually all elements of the periodic table. These advantages have made the SIMS measurements an indispensable tool for initial implant process setups, and process troubleshooting when process deviations are detected by TW, FPP or other monitoring techniques. the SIMS operation is based on sputtering the solid sample surface with a rastered 1–20 keV primary beam (usually oxygen or cesium, sometimes argon, gallium). Typically a 0.5×0.5 mm area is scanned. Ionized atoms sputtered from the scanned area are synchronously mass analyzed. The measured ion fluence for any particular mass plotted vs. sputter time yields a depth distribution of the analyzed species in the sample. Several variations of the SIMS techniques have evolved to provide optimum performance for particular tasks. The majority of SIMS tools use either magnetic sector, quadrupole, or time-of-flight mass spectrometers. Magnetic sector tools have the highest mass resolution and provide excellent

detection limits. Quadrupole mass spectrometers can rapidly switch from peak to peak, and are therefore well suited for the simultaneous acquisition of depth profiles of multiple elements. They also allow the achievement of improved depth resolution, when compared to SIMS magnetic sector tool. The need for accurate measurements of sub-50 nm depth profiles for ultra-shallow junctions, has stimulated development of SIMS measurements performed with reduced primary beam energy, high primary beam incidence angle, and oxygen stabilization of the sputtered crater surface (O-leak) [188]. The O-leak technique reduces the thickness of sample activated by the primary beam and stabilizes the sample surface during initial sputtering of the native oxide. As a result, improved accuracy and depth resolution of the first tens of nanometers from the sample surface are achieved. O-leak SIMS measurements were also demonstrated to provide superior performance for quantitative measurement of process induced surface contamination [190].

References

1. Bohr, N. *Kgl. Dan. Vid. Selks. Mat. Fys. Medd.* (1948): 18.
2. Lindhard, J., and M. Scharff. "Energy Dissipation by Ions in the keV Region." *Phys. Rev.* (1961): 124.
3. Lindhard, J., M. Scharff, and H. E. Schiott. *Kgl. Dan. Vid. Selks. Mat. Fys. Medd.* 33 (1963): 165.
4. Bethe, H. A. *Z. Phys.* 76 (1932): 293.
5. Bloch, F. *Ann. Phys.* 16 (1933): 287.
6. Lindhard, J., and A. Winther. *Kgl. Dan. Vid. Selks. Mat. Fys. Medd.* (1964): 34.
7. Biersack, J. P., and L. G. Haggmark. "A Monte Carlo Computer Program for the Transport of Energetic Ions in Amorphous Targets." *Nucl. Instrum. Methods* 174 (1980): 257.
8. Simionescu, A., G. Hobler, S. Bogen, L. Frey, and H. Ryssel. "Model for the Electronic Stopping of Channeled Ions in Silicon around the Stopping Power Maximum." *Nucl. Instrum. Phys. Res. B* 106 (1995): 47.
9. Brice, D. K. "Three-Parameter Formula for the Electronic Stopping Cross Section at Nonrelativistic Velocities." *Phys. Rev. A* 6 (1972): 1791.
10. Ziegler, J. F., J. P. Biersack, and U. Littmark. *The Stopping and Range of Ions in Solids*. New York: Pergamon Press, 1985.
11. Hobler, G., A. Simionescu, L. Palmetshofer, C. Tian, and G. Stingeder. "Boron Channeling Implantations in Silicon: Modeling of Electronic Stopping and Damage Accumulation." *J. Appl. Phys.* 77 (1995): 3697.
12. Firsov, O. B. *Sov. Phys. JETP* 36 (1959): 1076.
13. Oen, O. S., and M. T. Robinson. *Nucl. Instrum. Methods* 132 (1976): 647.
14. Klein, K. M., C. Park, and A. F. Tasch. "Local Electron Concentration-Dependent Electronic Stopping Power Model for Monte Carlo Simulation of Low-Energy Ion Implantation in Silicon." *Appl. Phys. Lett.* 57 (1990): 2701.
15. Chan, H. Y., K. Nordlund, J. Peltola, H.-J. L. Gossmann, N. L. Ma, M. P. Srinivasan, F. Benistant, and L. Chan. "The Effect of Interatomic Potential in Molecular Dynamics Simulation of Low Energy Ion Implantation." In *Proceedings of the 7th International Conference on Computer Simulation of Radiation Effects in Solids*, 42, 2004 (COSIRES 2004).
16. Gombas, P. *Die statistische Theorie des Atoms und ihre Anwendungen*. New York: Springer, 1949.
17. Ziegler, J. F., and J. P. Biersack. *SRIM-2003*, SRIM.com. Annapolis, MD, 2003, http://www.srim.org/
18. Heinrich, J. *A Guide to the Pearson Type IV Distribution*. University of Pennsylvania, http://www.cdf.fnal.gov/publications/cdf6820_pearson4.pdf, 2004.
19. Maes, H., W. Vandervorst, and R. van Overstraten. "Impurity Profile of Implanted Ions in Silicon." In *Impurity Doping Processes in Silicon*, edited by F. Y. Wang, 443. North-Holland, Amsterdam, Netherlands: Elsevier, 1981.
20. Gerhard Hobler. *IMSIL-2003.* Vienna, Austria: TU Vienna, 2003.
21. Tasch, A. F., and S. Banerjee. *UT-Marlowe*. Austin, TX: University of Texas, 1999.

22. Lindhard, J. "Influence of Crystal Lattice on Motion of Energetic Charged Particles." *Kgl. Dan. Vid. Selks. Mat. Fys. Medd.* 34 (1965): 1.
23. Barrett, J. H. "Monte Carlo Channeling Calculations." *Phys. Rev. B* 3 (1971): 1527.
24. Morgan, D. V. *Channeling*. New York: Wiley, 1973.
25. Gemmell, D. S. "Channeling and Related Effects in the Motion of Charged Particles through Crystals." *Rev. Mod. Phys.* 46 (1974): 129.
26. Feldman, L. C., J. W. Mayer, and S. T. Picraux. *Materials Analysis by Ion Channeling*. New York: Academic Press, 1982.
27. Firsov, O. B. *Sov. Phys. JETP* 6 (1957): 534.
28. Nielsen, O. H. *Phonons in Semiconductors in Internal Report*. Aarhus, Denmark: University of Aarhus, 1979.
29. Zieglerm, J. F., and R. F. Lever. "Channeling of Ions Near the Silicon ⟨001⟩ Axis." *Appl. Phys. Lett.* 46 (1985): 358.
30. Lever, R. F., and K. W. Brannon. "Crystallographic Aspects of Low Energy Boron Implantation into Silicon." *MRS Symp. Proc.* 100 (1988): 249.
31. Lever, R. F., and K. W. Brannon. "A Low Energy Limit to Boron Channeling in Silicon." *J. Appl. Phys.* 69 (1991): 6369.
32. Walther, S. R., et al. "Dopant Channeling as a Function of Implant Angle for Low Energy Applications." In *Proceedings of 12th International Conference on Ion Implant Technology*, edited by J. Matsuo, G. Takoaka, and I. Yamade, 126–9. Piscataway, NJ: IEEE Press, 1999.
33. Cristiano, F., B. Colombeau, and A. Claverie. "Energetics of Extrinsic Defects in Si and Their Role in Nonequilibrium Dopant Diffusion." *Defect and Diffusion Forum* 183 (2000): 199.
34. Knights, A. P., and F. Malik. "The Equivalence of Vacancy-Type Damage in Ion-Implanted Si Seen by Positron Annihilation Spectroscopy." *Appl. Phys. Lett.* 75 (1999): 466.
35. Pelaz, L., M. Jaraiz., et al. "B Diffusion and Clustering in Ion Implanted Si: The Role of B Cluster Precursors." *Appl. Phys. Lett.* 70 (1997): 2285.
36. Hobler, G., and G. Otto. "Status and Open Problems in Modeling of As-Implanted Damage in Silicon." *Mat. Sci. Semicond. Process.* 6 (2003): 1.
37. Semiconductor Industry Association. *The International Technology Roadmap for Semiconductors*. San Jose, CA: SIA, 2003 (http://public.itrs.net/Files/2003ITRS/Home2003.htm).
38. Kinchin, G. H., and R. S. Pease. "The Displacement of Atoms in Solids by Radiation." *Rep. Prog. Phys.* 18 (1955): 1.
39. Christel, L. A., J. F. Gibbons, and T. W. Sigmon. "Displacement Criterion for Amorphization of Silicon During Ion Implantation." *J. Appl. Phys.* 52 (1981): 7143.
40. Cerva, H., and G. Hobler. "Comparison of Transmission Electron Microscope Cross Sections of Amorphous Regions in Ion Implanted Silicon with Point-Defect Density Calculations." *J. Electrochem. Soc.* 139 (1992): 3631.
41. Hobler, G. In *Process Physics and Modeling in Semiconductor Technology*, edited by G. R. Srinivasan, C. S. Murthy, and S. T. Dunham, 509. Pennington, NJ: The Electrochemical Society, 1996.
42. Stein, H. J., et al. "Infrared Studies of the Crystallinity of Ion Implanted Si." *Radiat. Eff.* 6 (1970): 19.
43. Dennis, J. R., and E. B. Hale. "Crystalline to Amorphous Transformation in Ion-Implanted Silicon: A Composite Model." *J. Appl. Phys.* 49 (1978): 1119.
44. Maszara, W. P., and G. A. Rozgonyi. "Kinetics of Damage Production in Silicon During Self-Implantation." *J. Appl. Phys.* 60 (1986): 2310.
45. Claverie, A., C. Vieu, J. Fauré, and J. Beauvillain. "Cross-Sectional High-Resolution Electron Microscopy Investigation of Argon-Ion Implantation-Induced Amorphization of Silicon." *J. Appl. Phys.* 64 (1988): 4415.
46. Hobler, G. "Monte Carlo Simulation of Two-Dimensional Implanted Dopant Distributions at Mask Edges." *Nucl. Instrum. Methods Phys. Res. B* 96 (1995): 155.
47. Brown, R. A., et al. "Impurity Gettering to Secondary Defects Created by MeV Ion Implantation in Silicon." *J. Appl. Phys.* 84 (1998): 2459.

48. Venezia, V. C., et al. "Depth Profiling of Vacancy Clusters in MeV-Implanted Si Using Au Labeling." *Appl. Phys. Lett.* (1998): 73.
49. Kalyanaraman, R. "Quantitative Evolution of Vacancy-Type Defects in High-Energy Ion-Implanted Si: Au Labeling and the Vacancy Implanter." *Nucl. Instrum. Methods Phys. Res. B* 175–7 (2001): 182.
50. Goldberg, R. D., J. S. Williams, and R. G. Elliman. "Amorphization of Silicon by Elevated Temperature Ion Irradiation." *Nucl. Instrum. Methods Phys. Res. B* 106 (1995): 242.
51. Schultz, P. J., et al. "Crystalline-to-Amorphous Transition for Si-Ion Irradiation of Si(100)." *Phys. Rev. B* 44 (1991): 9118.
52. Pelaz, L. "Atomistic Modeling of Amorphization and Recrystallization in Silicon." *Appl. Phys. Lett.* 82 (2003): 2038.
53. Rubin, L., and R. Simonton. "Ion Implantation Applications in CMOS Process Technology." In *Ion Implantation: Science and Technology*, edited by J. F. Ziegler, 46. Yorktown, NY: Ion Implantation Technology Co., 2000.
54. Bourdelle, K. K., Y. Chen, R. A. Ashton, L. M. Rubin, A. Agarwal, and W. H. Morris. "Evaluation of High Dose, High Energy Boron Implantation into Cz Substrates for Epi-Replacement in CMOS Technology." *IEEE Trans. Electron Devices* 48 (2001): 2043.
55. Voldman, S., et al. "The Influence of Heavily Doped Buried Layer Implants on Electrostatic Discharge (ESD), Latchup, and a Silicon Germanium Heterojunction Bipolar Transistor in a BiCMOS SiGe Technology." *IEEE Int. Reliability Phys. Symp. Proc.* 25–9 (2004): 143.
56. Tsukamoto, K., et al. "High Energy Ion Implantation for ULSI: Well Engineering and Gettering." *Solid State Technol.* 35 (1992): 49.
57. Tamura, M., et al. "MeV Energy B+, P+, and As+ Ion Implantation into Si." *Nucl. Instrum. Methods Phys. Res. B* 21 (1987): 438.
58. Kuroi, T., et al. "Self-Gettering and Proximity Gettering for Buried Layer Formation by MeV Ion Implantation." *IEDM Tech. Dig.* (1990): 261.
59. Zappe, H., et al. "Characteristics of CMOS Devices in High Energy Boron Implanted Substrates." *IEEE Trans. Electron Devices* 35 (1988): 1029.
60. Erokhin, Y., et al. "Process Trends for Ultra-Low Energy and High Energy Ion Implantation." *Future Fab Int.* 1 (1997): 221.
61. Rubin, L., and W. Morris. "High Energy Ion Implanters and Applications Take Off." *Semicond. Int.* 20 (1997): 77.
62. Cheng, J. Y., et al. "Formation of Extended Defects in Silicon by High Energy Implantation of B and P." *J. Appl. Phys.* 80 (1996): 2105.
63. Wolf, S., *Silicon Processing for the VLSI Era*, Vol. 2, 5. Sunset Beach, CA: Lattice Press, 1990 chap. 5.
64. Hamada, A., et al. "N-Source/Drain Compensation Effects in Submicrometer LDD MOS Devices." *IEEE Electron Device Lett.* EDL-8 (1987): 398.
65. Krieger, G., et al. "Moderately Doped NMOS (M-LDD)-Hot Electron and Current Drive Optimization." *IEEE Trans. Electron Devices* 38 (1991): 121.
66. Yoshida, A., et al. "Hot Carrier Induced Degradation Mode Depending on the LDD Structure in MOSFETs." *IEDM Tech. Dig.* (1987): 42.
67. Chan, T. "Effects of the Gate-to-Drain/Source Overlap on MOSFET Characteristics." *IEEE Electron Device Lett.* EDL-8 (1987): 326.
68. Maitra, K., and N. Bhat. "Impact of Gate-to-Source/Drain Overlap Length on 80-nm CMOS Circuit Performance." *IEEE Trans. Electron Devices* 51 (2004): 409.
69. Capaletti, P., et al. "Application of Advanced Ion Implantation Techniques to Flash Memories." *Nucl. Instrum. Methods Phys. Res. B* 96 (1995): 405.
70. *International Technology Roadmap for Semiconductors*. San Jose, CA: Semiconductor Industry Association, 2003, available at http://public.itrs.net/
71. Agarwal, A., et al. "Boron-Enhanced Diffusion of Boron: The Limiting Factor for Ultra-Shallow Junctions." *IEDM Tech. Dig.* (1997): 367.
72. Gossmann, H. J., C. S. Rafferty, and P. Keys. "Junctions for Deep Sub-100 nm NMOS: How Far Will Ion Implant Take Us?" *Mater. Res. Soc. Symp. Proc.* 610 (2000): B1.2.1.

73. Agarwal, A., A. T. Fiory, H.-J. Gossmann, C. S. Rafferty, P. Frisella, and J. Hebb. "Ultra-Shallow Junction Formation by Spike Annealing in a Lamp-Based or Hot-Walled RTP System: Effect of Ramp-up Rate." *Mater. Sci. Semicond. Proc.* 1 (1999): 237.
74. Agarwal, A., et al. "Reduction of Transient Diffusion from 1–5 keV Si+ Ion Implantation Due to Surface Annihilation of Interstitials." *Appl. Phys. Lett.* 71 (1997): 3141.
75. Agarwal, A., H.-J. Gossmann, and A. T. Fiory. "Ultra-Shallow Junctions by Ion Implantation and Rapid Thermal Annealing: Spike-Anneals, Ramp Rate Effects." *Mater. Res. Soc. Symp. Proc.* 568 (1999): 19.
76. Giles, M. D. "Transient Phosphorus Diffusion below the Amorphization Threshold." *J. Electrochem. Soc.* 138 (1991): 138.
77. Stolk, P. A., et al. "Physical Mechanisms of Transient Enhanced Dopant Diffusion in Ion-Implanted Silicon." *J. Appl. Phys.* 81, no. 9 (1997): 6031.
78. Eaglesham, D. J., P. A. Stolk, H.-J. Gossmann, and J. M. Poate. "Implantation and Transient B Diffusion in Si: The Source of the Interstitials." *Appl. Phys. Lett.* 65 (1994): 2305.
79. Rafferty, C. S., G. H. Gilmer, M. Jaraiz, D. J. Eaglesham, and H.-J. Gossmann. "Simulation of Cluster Evaporation and Transient Enhanced Diffusion in Silicon." *Appl. Phys. Lett.* 68 (1996): 2395.
80. Gossmann, H.-J. "The Interstitial Fraction of Diffusivity of Common Dopants in Si." *Appl. Phys. Lett.* 71, no. 26 (1997): 3862.
81. Gossmann, H.-J. "Dopant and Point Defects during Silicon Processing." In *Semiconductor Silicon*, Vol. 98-1, edited by H. R. Huff, U. Goselle, and H. Tsuya, *ECS Proceeding*, 884, (1998).
82. Agarwal, A., H.-J. L. Gossmann, A. T. Fiory, V. C. Venezia, and D. C. Jacobson. "Ultra-Shallow Junction Formation Using Ion Implantation and Rapid Thermal Annealing: Physical and Practical Limits." *ECS Proc.* 2000-9, (2000).
83. Behrisch, R., ed. *Sputtering by Particle Bombardment*, Vol. 1–3. Berlin: Springer, 1983.
84. Agarwal, A. "Ultra-Shallow Junction Formation Using Conventional Ion Implantation and Rapid Thermal Annealing." In *2000 Conference on Ion Implantation Technology Proceedings*, edited by H. Ryssel, L. Frey, J. Gyulai, and H. Glawischnig, 293. Piscataway, NJ: IEEE Press, 2000.
85. Matsunami, N., Y. Yamamura, Y. Itakawa, N. Itoh, Y. Kazumata, S. Miyagawa, K. Morita, R. Shimizu, and H. Tawara. *At. Data Nucl. Data Tables* 31 (1984): 1.
86. Shockley, W. U.S. Patent 2787564.
87. McKenna, C. M. "A Personal Historical Perspective of Ion Implantation Equipment for Semiconductor Applications." In *2000 Conference on Ion Implantation Technology Proceedings*, edited by H. Ryssel, L. Frey, J. Gyulai, and H. Glawischnig, 1. Piscataway, NJ: IEEE Press, 2000.
88. Stephens, K. G. "Ion Source Physics." In *Ion Implantation Science and Technology*, Vol. 10598, edited by J. Ziegler, *Ion Implantation Science and Technology*, 465. Yorktown, NY: Ion Implantation Technology Co., 1996.
89. Freeman, J. "A New Ion Source for Electromagnetic Isotope Separators." *Nucl. Instrum. Methods* 22 (1963): 306.
90. White, N. "Ion Sources for Use in Ion Implantation." *Nucl. Instrum. Methods B* 37/38 (1989): 78.
91. Brown, R. L. "SDS™ Gas Source Feed Material Systems for Ion Implantation." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 311. Piscataway, NJ: IEEE Press, 1997.
92. Horsky, T., et al. "Performance and Lifetime of the Extended Life Ion Source." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 414. IEEE Press: Piscataway, NJ, 1997.
93. Curello, G., et al. "Characteristics of Extended Life Ion Source in Eaton's NV-8250 Medium Current Ion Implanter." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 223. Piscataway, NJ: IEEE Press, 1999.
94. Olsen, J., et al. "Varian Semiconductor Indirectly Heated Cathode Sources." In *2002 14th International Conference on Ion Implantation Technology Proceedings*, edited by B. Brown, et al., 417. Piscataway, NJ: IEEE Press, 2003.

95. Horsky, T. N., et al. "Current Status of the Extended Life Source: Lifetime and Performance Improvements." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 416. Piscataway, NJ: IEEE Press, 1999.
96. Jonoshita, I., M. Sugatini, and S. Takei. "ELS2:Extended Life Source with Dual Cathode." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 239. Piscataway, NJ: IEEE Press, 1999.
97. Farley, M., and B. Simonton. "Ion Source Operation and Maintenance." In *Ion Implantation Science and Technology*, edited by J. Ziegler, 511. Yorktown, NY: Ion Implantation Technology Co., 1996.
98. Humphries, S. *Charged Particle Beams*. 289–300. New York: Wiley, 1990.
99. Chang, B., et al. "Arsenic Dimer Implants for Shallow Extensions in 0.13 μm Devices." In *2002 14th International Conference on Ion Implantation Technology Proceedings*, edited by B. Brown, et al., 111. Piscataway, NJ: IEEE Press, 2003.
100. Jacobson, D. C., et al. "Decaborane, an Alternative Approach to Ultra Low Energy Ion Implantation." In *2000 Conference on Ion Implantation Technology Proceedings*, edited by H. Ryssel, L. Frey, J. Gyulai, and H. Glawischnig, 300. Piscataway, NJ: IEEE Press, 2001.
101. Perel, A., et al. "Decaborane Ion Implantation." In *2000 Conference on Ion Implantation Technology Proceedings*, edited by H. Ryssel, L. Frey, J. Gyulai, and H. Glawischnig, 304. Piscataway, NJ: IEEE Press, 2001.
102. Jacobson, D. C. private communication, SemEquip Co., 2005.
103. White, N., M. Sieradzki, and A. Renau. "The Ion Beam Optics of a Single Wafer High Current Ion Implanter." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 396. Piscataway, NJ: IEEE Press, 1997.
104. Graf, M. A., et al. "Low Energy Ion Beam Transport, Ion Implant Technology." In *2002 14th International Conference on Ion Implantation Technology Proceedings*, edited by B. Brown, et al., 359. Piscataway, NJ: IEEE Press, 2003.
105. Holmes, A. F. T. "Beam Transport." *Phys. Rev. A* 19 (1979): 389.
106. Sinclair, F., V. Benveniste, and J. Chen. U.S. Patent 5,814,819, Sept. 29, 1998.
107. Benveniste, V., W. F. DiVergilio, and F. Sinclair. U.S. Patent 6,414,329 B1 Jul. 2, 2002.
108. Benveniste, V., Y. Ye, and W. F. DiVergilio. U.S. Patent 6,541,781 B1 Apr. 1, 2003.
109. Benveniste, B., W. F. DiVergilio, and J. Z. Ye. U.S. Patent 6,759,665 B2 Jul. 6, 2004.
110. McIntyre, E. K., et al. "Purity of High Energy Beams in R.F. Linear Accelerator Basd Implanters." In *Proceedings 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 367. Piscataway, NJ: IEEE Press, 1997.
111. Wan, Z., K. Saadatmand, and F. Sinclair. "LINAC Simulation for High Energy Ion Implantation." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 371. Piscataway, NJ: IEEE Press, 1997.
112. Glavish, H. F. "Radio Frequency Linear Accelerators for Ion Implanters." *Nucl. Instrum. Methods B* 21 (1987): 218.
113. O'Connor, J. P., et al. "Performance Characteristics of the Genus Inc. Tandetron™ 1520 MeV Ion Implantation System." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 454. Piscataway, NJ: IEEE Press, 1997.
114. Tokoro, N., et al. "The Beam Performance of the Genus Tandetron 1520 MeV Implanter." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 443. Piscataway, NJ: IEEE Press, 1997.
115. Blake, J., and S. Richards. "The Ibis i2000 SIMOX Ion Implanter." In *2002 14th International Conference on Ion Implantation Technology Proceedings*, edited by B. Brown, et al., 391. Piscataway, NJ: IEEE Press, 2003.
116. Nagia, N., et al. "Nissin 300 mm Medium Current Ion Implanter EXCEED2300." In *2000 Conference on Ion Implantation Technology Proceedings*, edited by H. Ryssel, L. Frey, J. Gyulai, and H. Glawischnig, 415. Piscataway, NJ: IEEE Press, 2001.
117. Renau, A., and D. Hacker. "The VIISta 810 300 mm Medium Current Ion Implanter." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 158. Piscataway, NJ: IEEE Press, 1999.

118. Reimund, J., et al. "8250 Electrostatic Clamp Performance." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 280. Piscataway, NJ: IEEE Press, 1999.
119. Larson, K. "Improved Cooling on the Varian VIISta Series Ion Implanters." In *2000 Conference on Ion Implantation Technology Proceedings*, edited by H. Ryssel, L. Frey, J. Gyulai, and H. Glawischnig, 439. Piscataway, NJ: IEEE Press, 2001.
120. Mack, M. E. "Wafer Cooling and Wafer Charging in Ion Implantation." In *Ion Implantation Science and Technology*, edited by J. Ziegler, 538. Yorktown, NY: Ion Implantation Technology Co., 1996.
121. Macklin, R., et al. "Application of SEMI S8-95 for 200 and 300 mm Applied Materials Ion Implanters." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 146. Piscataway, NJ: IEEE Press, 1999.
122. Lundquist, P., et al. "The VIISion 80 and VIISion 200: High Current Ion Implantation Sysems for Greater Throughput with Excellent Performance at Low to High Doses." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 466. Piscataway, NJ: EEE Press, 1997.
123. Lee, H., et al. "Process Performance for Virtual Slot Disk/Triple Surface Disk Faraday on a Multi-Wafer High Current Ion Implant System." In *2002 14th International Conference on Ion Implantation Technology Proceedings*, edited by B. Brown, et al., 440. Piscataway, NJ: IEEE Press, 2003.
124. Mehta, S., et al. "Wafer Charge Neutralization Systems-Simulation and Experiment." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by IshidaE, et al., 57. Piscataway, NJ: IEEE Press, 1997.
125. Smatlak, D., M. Mack, and S. Mehta. "Charge Neutralization in Ion Implanters." *Nucl. Instrum. Methods B* 96 (1995): 22.
126. Current, M., M. Vella, and W. Lukaszek. "Beam-Plasma Concepts for Wafer Charging Control During Ion Implantation." In *Proceedings 11th International Conference on Ion Implant Technology*, edited by E. Ishida, et al., 53. Piscataway, NJ: IEEE Press, 1997.
127. Ito, H., et al. "High Density Plasma Flood System for Wafer Charge Neutralization." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 478. Piscataway, NJ: IEEE Press, 1999.
128. Sano, M., et al. "Plasma Electron Flood for a Scanned Beam Implanter." In *2002 14th International Conference on Ion Implantation Technology Proceedings*, edited by B. Brown, et al., 315. Piscataway, NJ: IEEE Press, 2003.
129. Smith, T. C. In *Photoresist and Particulate Problems*, Vol. 10598, edited by J. Ziegler, *Ion Implantation Science and Technology*, 629. Yorktown, NY: Ion Implantation Technology Co., 1996.
130. Chen, H. G., Y. Erokhin, E. McIntyre, F. Sinclair, and M. Sugitani. "Charge Exchange Effects in Production High Energy Ion Implanter Dosimetry." In *Proceedings of 11th International Conference on Ion Implant Technology*, edited by E. Ishida, et al., 104. Piscataway, NJ: IEEE Press, 1997.
131. Smatlak, D. L., J. Scheuer, A. Renau, A. Cucchetti, and J. Olson. "Beam Purity Control in VIISta 810 Implanter." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 166. Piscataway, NJ: IEEE Press, 1999.
132. Chang, B., J. Chang, A. Agarwal, M. Ameen, H. Chen, D. Chien, C. Tsai, C. Wang, D. Wu, and C. Yang. "Arsenic Dimer Implants For Shallow Extension in 0.13 Logic Devices." In *2002 14th International Conference on Ion Implantation Technology Proceedings*, edited by B. Brown, et al., 111. Piscataway, NJ: IEEE Press, 2003.
133. Kopalidis, P., C. Sohl, B. Freer, M. Ameen, R. Reece, and R. Rathmell. "Low Energy Implant Throughput Improvement by Using the Arsenic Dimer Ion on the Axcelis GSD III/LED Ion Implanter." In *2002 14th International Conference on Ion Implantation Technology Proceedings*, edited by B. Brown, et al., 122. Piscataway, NJ: IEEE Press, 2003.
134. *International Technology Roadmap for Semiconductors*, Front End Process Chapter 2003 Edition, http://public.itrs.net/2003itrs/home2003.htm
135. Lin, H., D.-Y. Lee, and T.-Y. Huang. "Breakdown Modes and Their Evolution in Ultrathin Gate Oxide." *Jpn. J. Appl. Phys.* 41 (2002): 5957.

136. Mack, M. E., and M. S. Ameen. "Wafer Cooling and Wafer Charging in Ion Implantation." In *Ion Implantation Science and Technology*, Vol. 10598, edited by J. F. Ziegler, 522. Yorktown, NY: Ion Implantation Technology Co., 2000.
137. Lukaszek, W. "Understanding and Controlling Wafer Charging Damage." *Solid State Technol.* 101, (1998).
138. Dixon, W., W. Lukaszek, and C. Heden. "Photoresist Enhanced Wafer Charging during High Current Ion Implantation." In *Proceedings 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 85. Piscataway, NJ: IEEE Press, 1997.
139. Holmes, A. F. T. "Beam Transport." *Radiat. Eff.* 44 (1978): 47.
140. Shin, H., and C. Hu. "Monitoring Plasma-Process Induced Damage in Thin Oxide." *IEEE Trans. Semicond. Manuf.* 6 (1993): 96.
141. Ma, S., and J. McVittie. "Real Time Measurement of Transients and Electrode Edge Effects for Plasma Charging Induced Damage." *IEDM Proc.* 94 (1994): 463.
142. Lukaszek, W., et al. "Characterization of Wafer Charging Mechanisms and Oxide Survival Prediction Methodology." *IEEE Int. Reliability Phys. Symp.* (1994): 334.
143. Current, M., W. Lukaszek, M. Vella, and N. Tripsas. "Surface Charge Control during High-Current Ion Implantation: Characterization with Charm-2 Sensors." *Nucl. Instrum. Methods Phys. Res. B* 96 (1995): 34.
144. Chan, Y. D. "Using SEMATECH Electrical Test Structures in Assessing Plasma Induced Damage in Poly Etching." *Jpn. J. Appl. Phys.* 33 (1994): 4458.
145. Heremans, P., J. Witters, G. Groeseneken, and H. Maes. "Analysis of the Charge Pumping Technique and Its Application for the Evaluation of MOSFET Degradation." *IEEE Electron Devices* 36 (1989): 1318.
146. Bauza, D., and Y. Maneglia. "In-Depth Exploration of Si–SiO_2 Interface Traps in MOS Transistors Using the Charge Pumping Technique." *IEEE Trans. Electron Devices* 44 (1997): 2262.
147. Morehead, F. F., and B. L. Crowder. "A Model for the Formation of Amorphous Si by Ion Implantation." In *1st International Conference on Ion Implantation*, edited by F. Eisen, and L. Chadderton. Thousand Oaks, NY: Gordon and Breach, 1971.
148. Tsai, J. C., and J. M. Marabito. "The Mechanism of Simultaneous Implantation and Sputtering by High Energy Oxygen Ions during Secondary Ion Mass Spectrometry (SIMS) Analysis." *Surf. Sci.* 44 (1974): 247.
149. Romig, T., M. Bishop, and V. Rio. "Exploration and Prevention of Photoresist Burning in High Current Ion Implanters." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 190. Piscataway, NJ: IEEE Press, 1997.
150. Fleming, D., M. Desarno, and R. Mohondro. "Manufacturing Improvements Realized through an Optimized Pre-Implant UV/Bake Process." *Future Fab Int.* 4 (1997): 177.
151. Vinogradova, G. K., and V. M. Menagarishvili. "Observations on the Formation and Ashing of Giant Folds in High Dose Ion-Implanted Resists." *J. Vac. Sci. Technol. B* 17 (1999): 95.
152. Fujimura, S., and H. Yano. *Nucl. Instrum. Methods B* 39 (1989): 809.
153. Orvek, K. J., and C. Huffman. "Carbonized Layer Formation in Ion Implanted Photoresist Masks." *Nucl. Instrum. Methods B* 7/8 (1985): 501.
154. Mohondro, R., et al. "Photostabilization." *Future Fab Int.* 3 (1996): 187.
155. Mohondro, R. "Photostabilization: Comparing DUV and i-Line." *Solid State Technol.* 46 (2003): 2.
156. Pittaway, L. G. "The Temperature Distributions in Thin Foil and Semi-Infinite Targets Bombarded by an Electron Beam." *Br. J. Appl. Phys.* 15 (1964): 967.
157. Adesida, I., C. Anderson, and E. D. Wolf. "Resist Exposure with Light Ions." *J. Vac. Sci. Technol. B* 1 (1983): 1182.
158. Jones, M. A., et al. "UV Curing and Photoresist Outgassing in High Energy Implantation." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 182. Piscataway, NJ: IEEE Press, 1997.
159. Perel, A. S., and T. N. Horsky. "Decay Rate of Photoresist Outgassing from Ion Implantation." *J. Vac. Sci. Technol. A* 18 (2000): 1800.

160. Smith, T. C. "Wafer Cooling and Photoresist Masking Problems." In *Ion Implantation Equipment and Techniques*, edited by H. Ryssel, and H. Glawischnig, 196. New York: Springer, 1983.
161. Fujimura, S., J. Konno, K. Hikazutani, and H. Yano. "Ashing of Ion-Implanted Resist Layer." *Jpn. J. Appl. Phys.* 28 (1989): 2130.
162. Axcelis Technologies Internal report FSD#1013, 2001.
163. Joshi, A., R. Nimmagadda, and J. Herrington. "Oxidation Kinetics of Diamond, Graphite, and Chemical Vapor Deposited Diamond Films by Thermal Gravimetry." *J. Vac. Sci. Technol. A* 8 (1990): 2137.
164. Tanaka, T., M. Yoshida, M. Shinohara, and T. Takagi. "Diamondlike Carbon Deposition on Plastic Films by Plasma Source Ion Implantation." *J. Vac. Sci. Technol. A* 20 (2002): 625.
165. Baba, K., et al. "Formation of Diamond Like Carbon Films by Plasma Source Ion Implantation from CH4, C2H2, and C6H6." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 1214. Piscataway, NJ: IEEE Press, 1999.
166. Smith, T. C. "Photoresist and Particulate Problems." In *Ion Implantation Science and Technology*, edited by J. Ziegler, 649. Yorktown, NY: Ion Implant Technology Co., 1996.
167. Gillespie, P., I. Berry, and P. Sakthivel. "Wafer Temperature Control—A Critical Parameter for Dry Photoresist and Residue Removal." *Semicond. Int.* October, (1999).
168. Bausum, T., M. DeSarno, and G. Dahrooge. "Stripping High-Dose Implanted Resist for 300 mm Production." *Semicond. Int.* June, (2003).
169. McOmber, J. I., K. Ostrowski, M. Meloni, R. Eddy, and P. Buccos. "Resist Preparation and Removal Techniques for High Dose Implantation on 200 mm Wafers." *Nucl. Instrum. Methods B* 74 (1993): 266.
170. Reinhardt, K., E. G. Pavel, N. Fernandes, D. Neil. "Development of an Integrated Solution for Total Wafer Cleaning." *IBM Technical Symposium*, France October 1999, oral presentation.
171. Kirkpatrick, A., N. Fernandes, T. Uk, and G. Patrizi. "Eliminating Heavily Implanted Resist in Sub-0.25-μm Devices." *MICRO* July/August (1998): 71.
172. Celler, G. K., D. L. Barr, and J. M. Rosamilia. "Thinnins of Si in SOI Wafers by SC-1 Standard Clean." In *SOI Conference Proceedings, IEEE*, 114, 1999.
173. Lui, L., K. L. Pey, and P. Foo. "HF Wet Etching of Oxide after Ion Implantation." In *Proceedings of Electron Devices Meeting*, 17. Hong Kong: IEEE, 1996.
174. The International Technology Roadmap for Semiconductors, 2003; http://public.itrs.net/Files/2003ITRS/Home2003.htm
175. Humphries, S. *Charged Particle Beams*. New York: Wiley, 1990 Chap. 3,5, Also available at: http://www.fieldp.com/cpb/cpb.html..
176. Ray, A. M., and J. P. Dykstra. "Beam Incidence Variations in Spinning Disk Ion Implanters." *Nucl. Instrum Method Phys. Res.* B55 (1991): 488.
177. Jones, M. A., and F. Sinclair. "Across-Wafer Channeling Variations on Batch Implanters: A Graphical Technique To Analyze Spinning Disk Systems." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 264. Piscataway, NJ: IEEE Press, 1997.
178. Stone, D., et al. "Performance of New Silicon Coated Disk Material: Disk Manufacturing Control & Device Production Experience." In *1998 International Conference on Ion Implantation Technology Proceedings*, edited by J. Matsuo, G. Takaoka, and I. Yamada, 574. Piscataway, NJ: IEEE Press, 1999.
179. Graff, K. In *Metal Impurities in Silicon Device Fabrication*, Vol. 24, edited by H.-J. Queisser, *Springer Series in Materials Science*, Berlin, New York, Heidelberg: Springer, 1995.
180. Wauk, M. T., et al. "Mechanism of Elemental Contamination in Ion Implantation Equipment." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 117. Piscataway, NJ: IEEE Press, 1997.
181. Current, M., L. A. Larson. "Ultra Pure Processing: A Key Challenge for Ion Implantation Processing for Fabrication of ULSI Devices." In *Proceedings of the Spring MRS Meeting*, edited by N. Chang, A. Marwick, and J. Roberto, 14, 365, 1989.
182. Heden, C., et al. "11BFH Contamination of Phosphorus Ion Implants." *J. Electrochem. Soc.* (1995).

183. Larson, L. A., M. I. Current, and C. Heady. "Enhanced Diffusion Effects and Dopant Cross Contamination in Ion Implanted Surfaces, Semiconductor Silicon." *Proc. Electrochem. Soc.* 86, no. 4 (1986): 667.
184. Natsuaki, N., et al. "ULSI Process Demands of Contamination Control on Ion Implantation." *Nucl. Instrum Method* B96 (1995): 62.
185. Smith, W. L., S. Rosencwaig, and D. L. Willenborg. "Ion Implant Monitoring with Thermal Wave Technology." *Appl. Phys. Lett.* 47 (1995): 584.
186. Martini, R., C. Wichard, and W. L. Smith. *Solid State Technol.* May, (1987).
187. Keenan, W. A., et al. "Advances in Sheet Resistance Measurements for Ion Implant Monitoring." *Solid State Technol.* June (1985): 155.
188. Eichinger, P. "Total Reflection X-Ray Fluorescence Analysis." In *Encyclopedia of Materials Characterization*, edited by C. R. Brundle, C. A. Evans, and S. Wilson, 361. Butterworth Heinemann, London, 1992.
189. Smith, S., L. Wang, J. Ericson, V. Chia, In *Proceedings of Materials Research Society Symposium*, 386, 1995.
190. Biswas, S., et al. "Analytical Techniques for Measuring Contamination Introduced during Ion Implantation." In *Proceedings of 11th International Conference on Ion Implantation Technology*, edited by E. Ishida, et al., 245. Piscataway, NJ: IEEE Press, 1997.

8

Dopant Diffusion

8.1 Introduction **8**-1
8.2 Definition of Point Defects **8**-3
Intrinsic Point Defects • Extrinsic Point Defects • Vacancy • Interstitial
8.3 Thermodynamics of Defects **8**-4
Enthalpy of Formation • Defect Concentrations at Thermal Equilibrium • Equilibrium between Different Change States of a Defect
8.4 Migration and Diffusion of Point Defects **8**-6
8.5 Fick's Laws of Diffusion **8**-7
8.6 Equilibrium Formulation for Dopant Diffusion **8**-9
Intrinsic Diffusion • Extrinsic Diffusion
8.7 Non-Equilibrium Formulation for Dopant Diffusion **8**-12
8.8 Diffusion in Strained Silicon **8**-14
8.9 Conclusions and Future Research **8**-16
Acknowledgments **8**-17
References **8**-17

Sanjay Banerjee
University of Texas at Austin

8.1 Introduction

Diffusion is a key area of ultra-large-scale-integration (ULSI) processing because, although dopants are generally introduced into a wafer by ion implantation, rather than thermally in a furnace, there is unavoidable diffusion of the dopants during any high-temperature process step. The ideas in this area can be categorized into two major approaches, namely; the continuum theory of Fick's diffusion equation and the atomistic theory [1,2]. The continuum theory requires the solution of Fick's diffusion equation with constant values for the diffusion coefficient and is adequate for low dopant concentrations. When the doping concentrations are high, the diffusion profiles exhibit anomalous diffusion behavior and a simple form of Fick's law cannot be applied because the diffusion coefficient becomes concentration dependent. The picture then becomes considerably more complicated and requires an atomistic approach, which studies the interactions between native point defects (vacancies and interstitials) and dopant atoms. The underlying idea behind all this is that the dopant atoms dissolve substitutionally in the lattice. Only through interactions with native point defects, the dopant atoms are able to jump from one site to another, effecting long-range diffusion.

This chapter will concern itself primarily with the atomistic theory of diffusion in silicon. The various types of point defects, followed by a description of the geometrical configuration of vacancies and interstitials are given in Section 8.2. The thermodynamics of different defect configurations and relations governing the concentrations of various defects will be derived and discussed in Section 8.3. In Section

8.4, the migration of the defects through the lattice will be discussed with emphasis on the dynamical theory. Fick's laws of diffusion will be discussed in Section 8.5. The formalism for intrinsic and extrinsic dopant diffusion under equilibrium and non-equilibrium conditions will be discussed in Section 8.6 and in Section 8.7, respectively.

Silicon-based heterostructures and strain are increasingly being exploited in the channel of metal oxide semiconductor field effect transistor (MOSFET) due to the enhanced mobility of the carriers. In Section 8.8, we discuss B diffusion in strained Si studied by using first principles density functional theory (DFT) calculations as an illustration of how strain can affect dopant diffusion. Silicon grown on a relaxed SiGe buffer layer is a common embodiment for generating biaxial tensile strain on the Si wafer. The biaxial tensile strain is induced in the thin epitaxial Si layer by the lattice mismatch between Si and SiGe. Presently, researchers have focused on not only better device performance that strained devices can achieve, but also other properties that are critical for process integration of strained materials. One of the most important properties is the dopant diffusion behavior under strain, because it impacts junction depth and effective channel length.

To motivate our discussions and to highlight the relevant issues for dopant diffusion, as it pertains to complementary metal oxide semiconductor (CMOS), we quote the following from the International Technology Roadmap for Semiconductors [3].

Advanced CMOS will probably evolve as follows as it pertains to doping technology.

Years 2005 through 2012—bulk silicon MOSFETS with the following enhancements with elevated contacts
Years 2008 through 2015—Fully-depleted SOI single gate planar devices with elevated contacts
Years 2011 through 2020—Fully-depleted, dual—or multi-gate devices, e.g., FINFET.

The difficult challenges for doping of CMOS transistors include achieving doping profiles in the source/drain extension regions to attain progressively shallower junction depths needed for control of short-channel effects ($\sim$10 nm), while concomitantly optimizing the sheet resistance ($\sim$500 Ohms/sq)-junction depth product, doping abruptness at the extension-channel junction, and extension-gate overlap; achieving controlled doping profiles in the channel region to set the threshold voltage, while concomitantly minimizing the short-channel effect and maximizing carrier mobility, and the formation of, and making low-resistance contact to shallow, highly doped source/drain regions. The implant energy and dose requirements as well as the resulting peak active dopant concentration are derived from the need to achieve an extension series resistance equal to 15% of the total series resistance, assuming dopant activation with negligible diffusion (i.e., flash or non-melt laser annealing or solid phase epitaxial regrowth). In a bulk planar MOSFET the as-implanted (vertical) junction depth with its proportional lateral straggle strongly influence subsequent lateral diffusion and encroachment of the channel. The short-channel behavior is therefore strongly linked to the vertical junction depth, and the drain extension resistance is strongly linked to doping concentration and lateral abruptness. The realization of ultra-shallow source and drain extension junction depths, that are vertically and laterally abrupt, requires not only the development of new and enhanced methods for implanting the doping species, but also the development of thermal activation processes that have an extremely small thermal budget. This is required to truncate the enhanced diffusion that accompanies the activation of the implanted dopant species.
For non-bulk, fully-depleted ultra-thin-body (FD-UTB) MOSFETs, doping processes require modification for optimized device drive current and threshold voltage stability. The critical extension junction depth is determined by the thickness of the active silicon layer; thus it becomes somewhat less challenging to make from an implant and anneal perspective. However, this does not imply that any implant energy is suitable for the extension of an UTB device, since the lateral junction is still linked to the (virtual) vertical one. The FD-UTB devices do not require channel doping to manage the short-channel effect, and therefore may be implemented using intrinsic and undoped silicon channels. However, the precise control of doping around the gate edge to optimize gate/drain overlap (or underlap) and the management of parasitic resistance remain important technology challenges.

8.2 Definition of Point Defects

A point defect in a crystal is an entity, which causes an interruption in the lattice periodicity [4]. Implicit in the definition is that the perturbation of the lattice is localized about a lattice site and involves only a few nearest neighbors. Point defects can be classified into four categories: intrinsic, extrinsic, associated or unassociated. *Intrinsic point defects* exist in the pure solid, while *extrinsic point defects* arise due to the introduction of impurity atoms into the lattice. Furthermore a point defect far enough away from another defect such that its properties remain unaffected is considered to be *isolated* or *unassociated*. If, however, it exists in a state such that it is interacting with another defect then it is considered to be *associated*.

8.2.1 Intrinsic Point Defects

Intrinsic point defects in semiconductors include vacancies, self-interstitials, and self-interstitialcies. An empty lattice site is called a *vacancy*, V, while a *self-interstitial*, I, is a host atom present on a site different from a regular substitutional lattice site. A *self-interstitialcy* is an associated defect comprising two atoms in non-substitutional positions configured about a single substitutional lattice site.

8.2.2 Extrinsic Point Defects

An impurity atom occupying a substitutional lattice site is known as a *substitutional defect*, A. If the impurity atom occupies an interstitial position, it is referred to as an *interstitial dopant*. When a vacancy is present next to a dopant atom, the defect is known as a *dopant–interstitialcy* AV pair. If an interstitial atom pairs with a substitutional dopant atom or if one of the atoms in an interstitialcy is a dopant (impurity) atom then the defect is called a dopant–interstitial/interstitialcy, respectively, but both are designated AI. In the discussion to follow later we will also use the symbol X to refer to I or V-type defects. In addition, interstitial-vacancy (or Frenkel pairs), divacancies etc., are also known to occur but they are not the favored forms for migration, and hence, will not be discussed in detail.

8.2.3 Vacancy

In order to form a vacancy by removing an atom from its lattice site, we have to break four bonds in a diamond structure [4]. The broken bonds can form new bonds depending on the charge state of the vacancy, which is just the number of electrons occupying the dangling bonds. This bonding causes small inward or outward displacements of neighboring atoms, which either preserve the local symmetry (relaxation) or alter it (distortion). The amplitude of these displacements depends on the charge state of these bonds. The detailed discussion of relaxation and distortions will be deferred to later sections. A split-vacancy configuration is one, where one neighbor of the vacancy is displaced midway between its original position and the center of the vacancy. This configuration is the saddle-point configuration of a normal vacancy during its migration.

8.2.4 Interstitial

The sites with high symmetry are stable interstitial positions for the interstitial atom, because in these positions the total electronic energy (with all atoms at their perfect lattice positions) will be at a minimum [4]. Several equivalent positions may exist in a unit cell due to the symmetry of the lattice and an interstitial atom can migrate between equivalent neighboring stable interstitial sites. It encounters other high-symmetry regions on the way assuming all remaining atoms are fixed at their undistorted lattice positions, corresponding to saddle points of electron energy.

These arguments, however, do not hold in the presence of phonons because then the lattice symmetry is altered. As a result the stable positions are no longer those with high symmetry but

rather "off-centered" configurations in which the interstitial is displaced slightly from its ideal site. In the diamond lattice the sites of highest symmetry are the hexagonal and tetrahedral sites. The "bond-centered" configuration is another position of high symmetry, a slight distortion of which leads to the interstitialcy configuration. Once again, the introduction of an interstitial causes relaxation and distortion of the surrounding lattice.

8.3 Thermodynamics of Defects

8.3.1 Enthalpy of Formation

A vacancy can be created by removing an atom from its substitutional site and placing it on the surface. The energy required to do this under constant pressure conditions is the enthalpy of formation of a vacancy, H_F^V. Similarly, the energy required to take an atom from the surface and place it on the chosen interstitial site is the enthalpy of formation of an interstitial, H_F^I. Hence, it can easily be seen that the formation enthalpy of a Frenkel pair (vacancy–interstitial pair) is $(H_F^V + H_F^I)$, assuming no interaction between the vacancy and the interstitial.

Two different theories have been developed, which determine formation enthalpies in covalent materials. The first theory is classical in nature [5,6] where the formation enthalpy is related to the bond dissociation energy, and an empirical potential (Morse potential) is applied to determine the variation of bond energy with inter-atomic distance. The second theory is a quantum-mechanical treatment based on the one-electron energy band approach [7]. The classical treatment will be briefly discussed in this section, since the quantum-mechanical treatment is considerably more complex. Specifically, the classical treatment will be discussed as it applies to the vacancy [5]. The formation of a vacancy requires the breaking of four bonds to remove a lattice atom and the forming of two new bonds to place this atom on the surface. New bonds are formed between the dangling bonds of the four neighbors of the vacancy, so that the neighbors undergo relaxation and distortion. Then

$$H_F^V = 2D - V_B - V_D \tag{8.1}$$

where D is the bond dissociation energy, V_B is the energy gained by bonding between the neighbors of the vacancy, and V_D is the energy associated with deformation of all other bonds. This method gives values of 1.97 and 2.35 eV for H_F^V in Ge and Si, respectively. For comparison, self-diffusion experiments in Ge and Si yielded values of 2.7 and 3.5 eV, respectively, for [8,9]. This discrepancy between values shows that the classical theory is greatly oversimplified.

8.3.2 Defect Concentrations at Thermal Equilibrium

In this section, the concentration of one kind of defect at low concentrations with no internal degrees of freedom will be discussed. The result will then be generalized for the case of several kinds of defects with internal degrees of freedom, and finally, the equilibrium between the different charge states of a defect will be considered.

8.3.2.1 One Kind of Defect at Low Concentrations

At a given temperature and pressure (generally 1 atm), the Gibbs free energy $G = H - TS$ is at a minimum. Since for solids, the volume changes during diffusion are negligible, it is a good approximation (as we will make here) to minimize the Helmholtz free energy instead. However, intrinsic defects are always present at any given temperature and their presence increases both the enthalpy H and the entropy S of the crystal. In the low concentration limit where each defect can be considered to be isolated, the defect concentration n can be obtained by minimizing the total change, ΔG, in free energy,

$$\Delta G = G - G_0 = n(H_F - TS_F) - TS_d, \tag{8.2}$$

where G_0 is the free energy of the perfect crystal, H_F is the enthalpy of formation of the defect, S_d is the configurational disorder over all possible lattice sites, and S_F is the disorder induced by lattice vibration with respect to n, leading to,

$$\frac{\partial \Delta G}{\partial n} = G_F - T\frac{\partial S_d}{\partial n} = G_F - T\frac{\partial k \ln W}{\partial n}, \tag{8.3}$$

where G_F is the free energy of formation of a single defect and W, the complexion number, is the number of the distinct ways to distribute n defects over N sites. Setting Equation 8.3 to zero in the low concentration limit gives the concentration of intrinsic defects in the crystal as

$$C_n = \frac{n}{N} = \exp\left(-\frac{G_F}{kT}\right). \tag{8.4}$$

8.3.2.2 Several Kinds of Defects with Internal Degrees of Freedom

There are generally several kinds of defects in real situations, and it is interesting to determine if Equation 8.4 can be generalized for a given type of defect. Defect potentials can generally be assumed to be short-range so that the defects can still be considered as independent. Therefore, the change in free energy, G, can be written as

$$\Delta G = \sum_i n_i G_{F_i} - kT \ln W, \tag{8.5}$$

where n_i is the number of defects of a particular kind, G_{F_i} is the free energy of this isolated defect, and W is the total complexion number, which has to be reevaluated for the present mixing condition. For the case of non-interacting defects on different lattice sites, W is given by

$$W = \frac{N!}{(N-n)!\prod_i n_i!}. \tag{8.6}$$

Following the earlier procedure, using Equation 8.5 and Equation 8.6 gives

$$C_{n_i} = \frac{n_i}{N} = \exp\left(-\frac{G_{F_i}}{kT}\right) \tag{8.7}$$

which generalizes the result given by Equation 8.4. In the case where a defect occurs with several internal degrees of freedom, Z_i, the above procedure can be repeated by simply replacing N by Z_iN, whence Equation 8.7 has to be replaced by

$$C_{n_i} = \frac{n_i}{N} = Z_i \exp\left(-\frac{G_{F_i}}{kT}\right). \tag{8.8}$$

8.3.3 Equilibrium between Different Charge States of a Defect

In semiconductors, defects can exhibit various charge states. For instance, in silicon, four charge states have been reported for the vacancy: V^+, V^0, V^-, and V^{--}. Under extrinsic conditions, not only does the relative population of the different charge states change depending on the position of the electro-chemical potential (Fermi level), but the total concentration of the point defects in the crystal also changes. This dependence is described below in the low concentration limit but with an overall constraint of global charge neutrality. It is possible to directly calculate the concentration of charged defects [10] and express it in terms of the Fermi level, considering reactions involving the charged defects, which

satisfy the charge neutrality constraint. For example, let us consider the reaction involving the ionization of the neutral vacancy V^0 into a positive vacancy, V^+ and an electron, e^-.

$$V^0 \leftrightarrow V^+ + e^-. \tag{8.9}$$

For this reaction, the free energy G is at a minimum at equilibrium with respect to n_0 and n_+, the number of neutral and positively charged vacancies, respectively. This implies that the equilibrium condition $\Delta G = 0$ can be written as

$$\frac{\partial G}{\partial n_0} = \frac{\partial G}{\partial n_+}, \tag{8.10}$$

which simply expresses the equality of chemical potentials. Solving Equation 8.10 allows the relative concentrations of the vacancies to be written as

$$\frac{C_{V^+}}{C_{V^0}} = \frac{Z_{V^+}}{Z_{V^0}} \exp\left[-\frac{G_F^{V^+} - G_F^{V^0} + E_F}{kT}\right] \tag{8.11}$$

where Z_{V^+} and Z_{V^0} are the internal degeneracies and E_F is the Fermi level. Equation 8.11 is completely general and allows the concentration of any charged vacancy to be obtained in terms of C_{V^0} and the Fermi level.

8.4 Migration and Diffusion of Point Defects

Atomistic models of diffusion in Si are based on the interaction of dopants with point defects in the silicon lattice. There are three native point defects of interest in Si—interstitials, vacancies, and interstitialcies. There exist various mechanisms that allow migration of defects within a lattice, depending on whether the defect is substitutional or interstitial. In interstitial diffusion, the interstitial moves from one interstitial site to an equivalent neighboring site without occupying a lattice site. The interstitial atom could also move by displacing a lattice atom, which, in turn, becomes an interstitial atom. This is an example of the interstitialcy mechanism. A related interstitialcy mechanism is the Crowdion mechanism, in which the interstitial atom located half-way between two lattice sites, migrates to one of the lattice sites and displaces the lattice atom. When a substitutional defect migrates by jumping from its original position to a neighboring vacancy site, the mechanism is called the vacancy mechanism. In general, the migration of any defect from one site to another requires the defect to jump over a barrier H_m. This jump probability is proportional to $\exp(-H_m/kT)$.

A defect migrates by jumping from one stable lattice site Q to a neighboring equivalent one R. There exists a potential energy barrier H_m^X to migration at the saddle-point between two stable sites. The frequency with which the defect surmounts the potential barrier requires knowledge of the dynamic interactions of the defect with the surrounding host atoms. Assuming a Boltzmann energy distribution, the rate of jumping can be intuitively written as,

$$\nu = \nu_D \exp\left(-\frac{H_m^X}{kT}\right) \tag{8.12}$$

where $\exp(-H_m^X/kT)$ is the probability that the defect will jump across the barrier and ν_D is the phonon (Debye) frequency. This activation law can be determined by using the dynamical theory [11], which treats the displacements causing migration as a superposition of phonons in the crystal. For example, a defect jumps when it has a sufficiently large amplitude of motion in the direction of the jump, and when the neighboring atoms at the saddle-point move sufficiently to reduce the closed shell

repulsion between them and the defect. According to this theory, the activation energy is given by,

$$H_{\mathrm{m}} = \frac{\pi^2 m_0 \nu_{\mathrm{D}}^2 a^2}{6} \tag{8.13}$$

where a is the separation of the neighboring atoms, or jump distance and m_0 is the mass of the migrating defect. This expression, however, does not take into account the displacements of neighboring atoms caused by the jump.

We next write the above general expressions more specifically for interstitial- and vacancy–assisted diffusion. In general, the diffusion coefficient is defined as

$$D = \frac{\nu a^2}{6} \tag{8.14}$$

where ν is the frequency of atomic jumping with the magnitude large enough to overcome the potential barrier. In the interstitial diffusion mechanism, the jumping frequency is given by [1].

$$\nu = 4\nu_{\mathrm{D}} \exp\left(-\frac{H_{\mathrm{m}}^{\mathrm{I}}}{kT}\right) \tag{8.15}$$

where $H_{\mathrm{m}}^{\mathrm{I}}$ is the potential barrier for the impurity migration from one interstitial site to another. Typical value for $H_{\mathrm{m}}^{\mathrm{I}}$ in Si are about 0.5–1.5 eV. Impurities, which belong to group I and VIII, such as lithium, potassium, sodium, argon, helium, and hydrogen, are known to move by this interstitial mechanism in Si.

In the vacancy mechanism, the probability that an adjacent substitutional site, which an atom can move into is vacant, must be taken into account. Hence,

$$\nu = 4\nu_{\mathrm{D}} \left(\frac{C_{\mathrm{V}}}{N_{\mathrm{Si}}}\right) \exp\left(-\frac{H_{\mathrm{m}}^{\mathrm{V}}}{kT}\right) \tag{8.16a}$$

where C_{v} is the total vacancy concentration, N_{Si} is the Si atomic density, and $H_{\mathrm{m}}^{\mathrm{V}}$ is the potential barrier for impurity migration from one substitutional site to another. At thermal equilibrium, the fraction of the lattice sites vacant in intrinsic Si is given by the equation

$$\frac{C_{\mathrm{V}}}{N_{\mathrm{Si}}} = \exp\left(-\frac{H_{\mathrm{f}}^{\mathrm{V}}}{kT}\right) \tag{8.16b}$$

where $H_{\mathrm{f}}^{\mathrm{V}}$ is the free energy for formation of a vacancy. Substituting Equation 8.16a and Equation 8.16b into Equation 8.14 yields

$$D = \left(\frac{4}{6}\right) \nu_{\mathrm{D}} a^2 \exp\left\{-\frac{\left(H_{\mathrm{m}}^{\mathrm{V}} + H_{\mathrm{f}}^{\mathrm{V}}\right)}{kT}\right\} \tag{8.17}$$

The activation energy for substitutional diffusion ($=H_{\mathrm{m}}^{\mathrm{V}} + H_{\mathrm{f}}^{\mathrm{V}}$) is larger than that for interstitial diffusion ($=E_{\mathrm{m}i}$) and is usually between 3 and 5 eV in Si. For the substitutional impurities (Group III and V, for example, As, P, and B) in Si, the activation energy for diffusion is larger than 3 eV.

8.5 Fick's Laws of Diffusion

Due to the agitation of the lattice by phonons some of the defects can wander throughout the lattice. For a simple cubic lattice this diffusion of defects can be understood by considering the jump process between

two adjacent (100) planes, 1 and 2. If the lattice planes contain n_1 and n_2 defects per unit surface area, respectively, and the jump rate in either direction is given by ν then the number of defects per unit surface area jumping from plane 1 to 2 in time dt is $J_1 = n_1 \nu \mathrm{d}t$. For the same jump probability in either direction, the net flux of dopants from plane 1 to 2, J_{12}, can be written as [1]

$$J_{12} = \frac{J_1 - J_2}{\mathrm{d}t} = (n_1 - n_2)\nu \tag{8.18}$$

For a small lattice constant a and assuming that the number of defects changes slowly with distance x, the above flux can be written as

$$J = -\nu a(\partial n/\partial x) \tag{8.19}$$

Since the defect concentration per unit volume $C = n/a$, Equation 8.19 becomes

$$J = -a^2\nu(\partial C/\partial x) = -D(\partial C/\partial x) \tag{8.20}$$

where $D = a^2\nu$ is the diffusion coefficient or diffusivity. The above equation is Fick's first law of diffusion. From the expression for ν given in Equation 8.12, it is seen that the diffusion coefficient is thermally activated.

$$D = D_0 \exp\left(-\frac{Q}{kT}\right) \tag{8.21}$$

where Q is the activation energy. Q is just H_m as can be observed by comparing with Equation 8.12 for ν, and D_0 is proportional to the Debye frequency. Thus, the activation energy for self-diffusion by the interstitial mechanism is just H_m^I, while for the vacancy mechanism, it is given by the sum $(H_\mathrm{m}^\mathrm{V} + H_\mathrm{f}^\mathrm{V})$.

We now consider the case, where the impurity concentration and its gradient vary with annealing time. From the continuity equation (conservation of matter), the net decrease in impurity concentration per unit time is determined by the divergence of the diffusion flux.

$$\frac{\partial C(x,t)}{\partial t} = -\frac{\partial J(x,t)}{\partial x} \tag{8.22}$$

Substituting Equation 8.20 into Equation 8.22, we obtain Fick's Second Law.

$$\frac{\partial C(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial C(x,t)}{\partial x}\right) \tag{8.23}$$

In the presence of an electric field, however, this equation is modified to

$$\frac{\partial C(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial C(x,t)}{\partial x}\right) \pm \frac{q}{\mathrm{kT}}\frac{\partial}{\partial x}\left(\mathrm{DC_e}\frac{\partial \Phi}{\partial x}\right) \tag{8.24}$$

where C_e is the electrically charged impurity concentration and Φ is the electrostatic potential, which is given by

$$\Phi = \frac{\mathrm{kT}}{q}\ln\frac{n}{n_\mathrm{i}} \tag{8.25}$$

where n and n_i are the extrinsic and intrinsic carrier concentrations, respectively.

If the impurity concentration is lower than n_i at the diffusion temperature, the diffusion coefficient is found to be independent of the impurity concentration, and hence solving Equation 8.24 is straightforward. For most of diffusion processes in ULSI device fabrication, however, the dopant concentration is very high, such that as discussed below, charged vacancies play a dominant role and dependence of diffusivities on the impurity concentration must be taken into account. It should be noted that the impurity concentration is a function of time and position. In this case, it is often impossible to obtain solutions to Equation 8.24 in a closed form.

8.6 Equilibrium Formulation for Dopant Diffusion

Although below solid solubility most of the dopant atoms dissolve on substitutional sites, at any given temperature, a finite number of dopant–defect pairs will be present. This fraction is responsible for the redistribution of dopant atoms, i.e., defects can migrate in the AV, AI, and A_i states. For the following discussion, diffusion is regarded as quasi-equilibrium process with a mass action relationship between A, X, and AX defects, although by nature diffusion is a non-equilibrium process. It is not necessary to specify the type of defect species (I or V) involved under quasi-equilibrium conditions, which in turn, implies that experiments performed under these conditions cannot distinguish between the relative contributions of interstitials or vacancies. Following Fahey et al's approach [12], we will not distinguish between the interstitial and interstitialcy mechanisms, and both mechanisms will be included in the representation AI.

8.6.1 Intrinsic Diffusion

For a dopant, A, which diffuses by either I- or V-type mechanisms, the dopant flux can be described by Fick's law as

$$J_A = -\left(d_{AV}\frac{\partial C_{AV}}{\partial x} + d_{AI}\frac{\partial C_{AI}}{\partial x}\right) \tag{8.26}$$

where d_{AV} and d_{AI} are the diffusivities of AV and AI, respectively. To re-express J_A in terms of a measurable quantity, C_A, a relationship has to be found between C_A, C_{AI}, and C_{AV}. This can be done by considering chemical reactions of the type

$$A + X \Leftrightarrow AX. \tag{8.27}$$

Under quasi-equilibrium conditions, local equilibrium exists at each point along the doping profile, and therefore the forward reaction rate K can be expressed as $K(T) = (C_A C_X / C_{AX})$, where K depends only on temperature, T. From the relation for K under equilibrium conditions, C_{AX}/C_A should be constant, since C_X is independent of the doping concentration under equilibrium conditions. Therefore, we can write

$$\frac{\partial C_{AX}}{\partial x} = \frac{C_{AX}}{C_A}\frac{\partial C_A}{\partial x}, \tag{8.28}$$

and Equation 8.26 can be rewritten in terms of C_A as

$$J_A = -\left[d_{AV}\left(\frac{C_{AV}}{C_A}\right)^* + d_{AI}\left(\frac{C_{AI}}{C_A}\right)^*\right]\frac{\partial C_A}{\partial x} = -(D^*_{AV} + D^*_{AI})\frac{\partial C_A}{\partial x}. \tag{8.29}$$

Since present experimental techniques can only measure macroscopic diffusivity, i.e.,

$$D^*_A = D^*_{AV} + D^*_{AI}, \tag{8.30}$$

it is not possible to measure D^*_{AV} and D^*_{AI} individually. This diffusivity is a function of temperature alone, and has a thermal activation energy Q_A, as shown earlier in Equation 8.21. The activation energy for self-diffusion, Q_{self}, in silicon has been found to be ~5 eV [13]. For dopant diffusion, however, the activation energy Q_A is found to be ~1 eV less than that for Si self-diffusion [14]. This difference in activation energy, Q_A, depends on the dopant species and the diffusion mechanism. Due to the immense complexity of the subject, and also because it is extremely hard to verify the theoretical estimates from the experimental activation energy data without knowledge of the dominant diffusion mechanism, this will not be discussed further.

8.6.2 Extrinsic Diffusion

Under quasi-equilibrium conditions, it is not necessary to know the details of the defect reactions even for extrinsic doping in order to model the diffusion accurately. For extrinsic doping, however, unlike in the case of intrinsic doping the Fermi level position in the bandgap varies spatially over the doping profile since the carrier concentration changes with the ionized dopant concentration. This leads to concentration-dependent diffusivities since the equilibrium ratio of $C_A C_X / C_{AX}$ also changes spatially with the Fermi level. The spatial variation of the Fermi level also results in a built-in electric field over the extent of the diffusion profile, which results in an enhancement of diffusion. For a single species diffusing under extrinsic doping conditions the doping profile can be obtained by solving

$$\frac{\partial C_A}{\partial t} = \frac{\partial}{\partial x}\left[D^*_A \frac{\partial C_A}{\partial x}\right] \tag{8.31}$$

where

$$D^*_A = h\left[D^i_{AX^0} + D^i_{AX^+}\left(\frac{p}{n_i}\right) + D^i_{AX^-}\left(\frac{n}{n_i}\right) + D^i_{AX^=}\left(\frac{n}{n_i}\right)^2\right] \tag{8.32}$$

and the electric field factor h is defined as

$$h = 1 + \frac{C_A}{2n_i}\left[\left(\frac{C_A}{2n_i}\right)^2 + 1\right]^{-1/2}. \tag{8.33}$$

The electric field factor typically varies between 1 and 2 for $n \ll n_i$ and $n \gg n_i$, respectively. The superscript "i" denotes intrinsic doping conditions. Equation 8.32 is a generalized form for dopant diffusivity, which includes all major combinations of dopant–defect interactions. Diffusion of impurities in silicon, as mentioned earlier, takes place through a combination of interstitial and vacancy mechanisms. Although the relative importance of the two is not clear in some cases, it is believed that both mechanisms play an important role in impurity diffusion since both I- and V-type defects are present in Si at the processing temperatures.

To explain Si self-diffusion and diffusion of group III and V impurities in Si, diffusion models based on the vacancy mechanism have been developed during the last three decades and have been used in process simulation programs such as SUPREM4 for modeling diffusion under non-oxidizing conditions [15]. The theory behind the vacancy mechanism is based on the interaction of impurities with vacancies. With the assumption that vacancies with different charge states interact with impurities independently, the diffusion coefficient in extrinsic silicon can be described by

$$D = D^0 + D^+ + D^- + D^= \tag{8.34}$$

where D^0, D^+, D^-, $D^=$ are the diffusion coefficients associated with the interaction of impurities with neutral vacancies, singly-positively charged vacancies, singly-negatively charged vacancies, and doubly-

negatively charged vacancies, respectively. The diffusion coefficient is proportional to the vacancy concentration. Therefore, Equation 8.34 can be expressed as

$$D = D_{\mathrm{i}}^{0}\left[\frac{C_{\mathrm{v}_0}}{C_{\mathrm{v}_0\mathrm{i}}}\right] + D_{\mathrm{i}}^{+}\left[\frac{C_{\mathrm{v}+}}{C_{\mathrm{v}+\mathrm{i}}}\right] + D_{\mathrm{i}}^{-}\left[\frac{C_{\mathrm{v}-}}{C_{\mathrm{v}-\mathrm{i}}}\right] + D_{\mathrm{i}}^{=}\left[\frac{C_{\mathrm{v}=}}{C_{\mathrm{v}=\mathrm{i}}}\right] \tag{8.35}$$

where $D_{\mathrm{i}}^{\mathrm{r}}$ are diffusivities in intrinsic Si due to vacancies in charge state "r", and C_{vr} and C_{vri} are the vacancy concentrations in charge state "r" in extrinsic Si and in intrinsic Si, respectively. Using the law of mass action at thermal equilibrium, it can be shown that

$$\frac{C_{\mathrm{vo}}}{C_{\mathrm{voi}}} = 1, \quad \frac{C_{\mathrm{v}+}}{C_{\mathrm{v}+\mathrm{i}}} = \frac{p}{n_{\mathrm{i}}}, \quad \frac{C_{\mathrm{v}-}}{C_{\mathrm{v}-\mathrm{i}}} = \frac{n}{n_{\mathrm{i}}}, \quad \frac{C_{\mathrm{v}=}}{C_{\mathrm{v}=\mathrm{i}}} = \left[\frac{n}{n_{\mathrm{i}}}\right]^2 \tag{8.36}$$

where n and p are the free electron and hole concentrations. Therefore, Equation 8.36 becomes, under extrinsic conditions

$$D = D_{\mathrm{i}}^{0} + D_{\mathrm{i}}^{+}\left(\frac{p}{n_{\mathrm{i}}}\right) + D_{\mathrm{i}}^{-}\left(\frac{n}{n_{\mathrm{i}}}\right) + D_{\mathrm{i}}^{=}\left(\frac{n}{n_{\mathrm{i}}}\right)^2 \tag{8.37}$$

For the diffusion of group III and V impurities of interest, all of the components associated with various charged states are not significant. From iso-concentration studies [16,17] it has been found that specific charge states dominate various common dopants. For acceptors such as B which are negatively charged, the neutral and positively charged vacancies contribute to diffusion, while the positively charged donors, such as As and P diffuse with neutral and negatively charged vacancies.

Thus, the D^0 and D^+ components are dominant for B diffusion, so that B diffusivity can be described by [15]

$$D(\mathrm{B}) = D_{\mathrm{i}}^{0} + D_{\mathrm{i}}^{+}(p/n_{\mathrm{i}}) = \left[0.037 + 0.72(p/n_{\mathrm{i}})\right]\mathrm{e}^{-3.46\ \mathrm{eV}/kT}\mathrm{cm}^2/\mathrm{s} \tag{8.38}$$

Similarly, for As diffusion [15]

$$D(\mathrm{As}) = D_{\mathrm{i}}^{0} + D_{\mathrm{i}}^{-}(n/n_{\mathrm{i}}) = 0.066\ \mathrm{e}^{-3.44\ \mathrm{eV}/kT} + 12.0(n/n_{\mathrm{i}})\mathrm{e}^{-4.05\ \mathrm{eV}/kT\ \mathrm{cm}^2/\mathrm{s}} \tag{8.39a}$$

The chemical concentration of As, C_{T}, is related to the free electron concentration by

$$C_{\mathrm{T}}(\mathrm{As}) \cong n + 1.13 \times 10^{-68}\mathrm{e}^{+1.60\ \mathrm{eV}/kT}n^4 \tag{8.39b}$$

where C_{T} is the chemical concentration of As.

For P diffusion [15]

$$\begin{aligned} D(\mathrm{P}) &= D_{\mathrm{i}}^{0} + D_{\mathrm{i}}^{-}(n/n_{\mathrm{i}}) + D_{\mathrm{i}}^{=}(n/n_{\mathrm{i}})^2 \\ &= 3.85\ \mathrm{e}^{-3.66\ \mathrm{eV}/kT} + 4.44(n/n_{\mathrm{i}})\mathrm{e}^{-4.00\ \mathrm{eV}/kT} + 44.2(n/n_{\mathrm{i}})^2\mathrm{e}^{-4.37\ \mathrm{eV}/kT}\mathrm{cm}^2/\mathrm{s} \end{aligned} \tag{8.40}$$

$$C_{\mathrm{T}}(\mathrm{P}) \cong n + 5.33 \times 10^{-43}\mathrm{e}^{+0.40\ \mathrm{eV}/kT}n^3 \tag{8.41}$$

For P diffusion, however, there are some deviations from the simple theory. The shape of high concentration P diffusion profiles in Si is characterized by a relatively flat high concentration region, a "kink" in the P profile, and an extensive tail region where the P diffusivity is enhanced by more than two orders of magnitude. To explain the enhancement of P diffusivity in the tail region, Fair and Tsai have developed a model which is based on the dissociation of the surface $\mathrm{P}^{+}\mathrm{V}^{=}$ pairs [18]. When the

concentration falls below a critical concentration (n_e), the Fermi level moves past the $V^=$ energy level (0.11 eV below the conduction band edge), and the $P^+V^=$ pairs dissociate to P^+V^- pairs, releasing free electrons, thereby resulting in a drastic increase of the P^+V^- pair concentration. This eventually increases the V^- concentration in the tail region because the binding energy of the P^+V^- pair is 0.3 eV less than for the $P^+V^=$ pair, and thus the P^+V^- pair prefers to be further dissociated to P^+ and V^-. The P diffusivity in the tail region is thus given by

$$D = D_i^o = D_i^- \left[\frac{C_{v+}}{C_{v+i}}\right] = 3.85\, e^{-3.66\ eV/kT} + 4.44\left(\frac{n}{n_i}\right) e^{-4.00\ eV/kT}\left[\frac{n_s^2}{n_e^2 n_i}\left(1 + e^{+0.3\ eV/kT}\right)\right] cm^2/s \tag{8.42}$$

where n_s is the surface electron concentration.

The vacancy mechanism has been well developed and agrees reasonably well with experimental results in some cases for diffusion of group III and V impurities (e.g., As, P, B, and Sb) in Si. However, there are still many issues in diffusion, which cannot be explained by a "vacancy-only" mechanism; it has serious shortcomings and cannot explain a lot of diffusion phenomena unless the interstitial contribution is also included. For example, the "vacancy-only" mechanism cannot explain the "emitter-push" effect [18], high concentration P diffusion [19], anomalous transient-enhanced diffusion in single crystal Si [20,21] and oxidation-enhanced or retarded diffusion (OED or ORD) [22–24]. In n–p–n BJTs with heavily phosphorus-doped emitters, the boron-doped base underneath the emitter shows enhanced diffusion relative to base diffusion outside the emitter. This was earlier attributed to supersaturation of vacancies in the tail region of the high concentration P profile resulting from the dissociation of the $P^+V^=$ pair. However, it is now believed [19] that high concentration P diffusion results in a supersaturation of interstitials and undersaturation of vacancies in the tail region of the profile. This suggests that interstitials also significantly contribute to B diffusion since the undersaturation of vacancies would result in retarded diffusion if the "vacancy-only" mechanism was employed to explain B diffusion. It is known from the growth and shrinkage of extrinsic stacking faults that oxidation results in a supersaturation of interstitials and undersaturation of vacancies in the Si substrate [25,26]. This approach, with certain assumptions, has led to the conclusion that some dopants, such as P and B diffuse predominantly by a interstitial mechanism, some such as Sb diffuse by a vacancy mechanism, while others such as As diffuse with both interstitial and vacancy components or a "dual-diffusion" mechanism. In the next section, we will discuss the formulation for dopant diffusion under non-equilibrium conditions.

8.7 Non-Equilibrium Formulation for Dopant Diffusion

In the last twenty years, a new diffusion model has been developed and the results of the model show good agreement with experiment [27–31]. One of the two key differences between the new model and the previous models is that a dual-diffusion mechanism is invoked in the new model. In the dual-diffusion mechanism, substitutional impurities are assumed to move both via vacancy-assisted diffusion and via interstitial-assisted diffusion. The vacancy-assisted diffusion mechanism is a general formulation, and therefore it can represent the vacancy diffusion mechanism or the impurity–vacancy pair diffusion mechanism [37]. Similarly, the interstitial-assisted mechanism can also represent the interstitialcy mechanism [33] or the impurity–interstitial pair diffusion mechanism [32].

The second one of the two key differences between the new and old models is that the new model is based on non-equilibrium concentrations of point defects [27–32]. For non-equilibrium point defect concentrations, the coupled equations for impurities and point defects must be solved because impurities continue to have dynamic interactions with point defects during annealing. The fact that point defect

concentrations are in non-equilibrium during diffusion can be easily proven from experimental results. For example, different types of crystal damage due to implantation continue to evolve to generate a number of point defects during annealing, resulting in anomalous transient diffusion in the substrate [34]. These types of crystal damage include point defect clusters, amorphous layers, projected range dislocations and end-of-range dislocations, depending on implant species, dose, energy, wafer temperature, and orientation. Some chemical interactions occurring at the surface also generate point defects, for example, interstitials by oxidation and vacancies by nitridation, thereby enhancing or retarding impurity diffusivities in the substrate [35–37]. It is also believed that the "emitter-push" effect is caused by injection of interstitials from a heavily P-doped region into the substrate. All these results consistently show that point defect concentrations are at non-equilibrium values during annealing.

Non-equilibrium point defect concentrations are formed during certain thermal processing steps during device fabrication such as oxidation or nitridation, which can cause enhanced or retarded diffusion as mentioned earlier. Sometimes, even thermal annealing of ion implantation in an inert ambient results in transient enhancement of the point defect concentrations. The "dual-diffusion" mechanism has been employed to explain the observed results in these experiments. It has generally been assumed that all substitutional dopants diffuse via a "dual-diffusion" mechanism with a fraction, f_I, of the dopants diffusing via the interstitial mechanism and a fraction, $f_V = 1 - f_I$, diffusing via the vacancy mechanism. The time-averaged diffusivity $\langle D \rangle$ under intrinsic doping conditions is given by

$$\frac{\langle D \rangle}{D^i} = \left[f_I \frac{C_I}{C_I^*} + f_V \frac{C_V}{C_V^*} \right], \quad \text{where } f_I = \frac{D_{AI}^*}{D_{AI}^* + D_{AV}^*} \tag{8.43}$$

In the above relation for diffusivity, C_V and C_I are the vacancy and interstitial concentrations, respectively, and the asterisks refer to quantities under equilibrium conditions. Orlowski [38] has recently shown that it is necessary to solve for the point defect concentrations explicitly for diffusion under non-equilibrium conditions. The "dopant–point defect pair" diffusion models [39–46] comprehend this by simultaneously solving for interstitial and vacancy concentrations along with the doping concentration.

In the pair diffusion model, it is assumed that the dopants diffuse in the form of dopant–interstitial or dopant–vacancy pairs. This mechanism was first described by Morehead and Lever [27] to explain the anomalous enhanced tail for high-concentration boron and phosphorus diffusion. According to their model, the dopant combines with the interstitial at the surface and diffuses into the interior in the form of a dopant–interstitial pair. The interstitial is released when the dopant becomes substitutional, which may either diffuse into the bulk or towards the surface. This model assumes that the flux of dopant–interstitial pairs into the bulk at any point is balanced by the flux of interstitials towards the surface. Mulvaney and Richardson [28] relaxed this condition of local equality of fluxes in order to include dynamic effects. The diffusion equations are derived by considering chemical reactions of the type given in Equation 8.27 among the dopants, point defects and free carriers. A generalized form for dopants can be obtained by substituting concentrations of the various dopant–defect complexes in Fick's second law.

$$\begin{aligned} \frac{\partial C_A}{\partial t} = {} & \frac{\partial}{\partial x} \left[\frac{f_I D_A^i}{C_I^*} \frac{\partial (C_I C)}{\partial x} + Z \frac{f_I D_A^i}{C_I^*} C_I C \frac{\partial \ln(n)}{\partial x} \right] \\ & + \frac{\partial}{\partial x} \left[\frac{(1 - f_I) D_A^i}{C_V^*} \frac{\partial (C_V C)}{\partial x} + Z \frac{(1 - f_I) D_A^i}{C_V^*} C_V C \frac{\partial \ln(n)}{\partial x} \right], \end{aligned} \tag{8.44}$$

where C_A is the total dopant concentration, C_I is the interstitial concentration, C_I^* is the equilibrium concentration of interstitials, C_V is the vacancy concentration, C_V^* is the equilibrium concentration of vacancies, Z is the charge state of the dopant atoms ($+1$ for donors; -1 for acceptors), f_I is the fractional contribution of the interstitial mechanism for impurity diffusion, and D_A^i is the intrinsic dopant diffusivity. The log terms in the above equation are the fluxes due to the built-in electric field.

The equation for the interstitial and vacancy concentrations can be obtained similarly. These equations include an additional term proportional to $(C_I C_V - C_I^* C_V^*)$ since a reaction of the form

$$I + V = \langle 0 \rangle \tag{8.45}$$

describing the generation-recombination process must be included when both interstitials and vacancies are present. The full equations for the interstitial and vacancy concentrations are

$$\frac{\partial C_I}{\partial t} = \frac{\partial}{\partial x}(D_I \frac{\partial C_I}{\partial x}) - K_R(C_I C_V - C_I^* C_V^*) + \frac{\partial}{\partial x}\left[\frac{f_I D_A^i}{C_I^*}\frac{\partial (C_I C)}{\partial x} + Z\frac{f_I D_A^i}{C_I^*} C_I C \frac{\partial \ln(n)}{\partial x}\right], \tag{8.46}$$

and

$$\frac{\partial C_V}{\partial t} = \frac{\partial}{\partial x}(D_V \frac{\partial C_V}{\partial x}) - K_R(C_I C_V - C_I^* C_V^*) + \frac{\partial}{\partial x}\left[\frac{1 - f_I D_A^i}{C_V^*}\frac{\partial (C_V C)}{\partial x} + Z\frac{(1 - f_I) D_A^i}{C_V^*} C_V C \frac{\partial \ln(n)}{\partial x}\right]. \tag{8.47}$$

Here, D_I and D_V are the interstitial and vacancy diffusivities and K_R is the defect annihilation rate. Equation 8.44, Equation 8.46, and Equation 8.47 essentially constitute the dopant diffusion model contained in the simulation programs PEPPER or FLOOPS [32]. It is interesting to note that in the limit $C_I = C_I^*$ and $C_V = C_V^*$, Equation 8.40 reduces to the standard vacancy model given in Equation 8.31 along with a term for the flux due to the built-in electric field. However, if C_I and C_V are constant but not equal to their equilibrium concentrations, Equation 8.40 reduces to the standard model with an effective diffusivity given by Equation 8.43.

One of the key parameters in the model is f_I and its values have been determined by fitting the experimental data to the model. The values of f_I used in PEPPER are given by [32]

$$f_I = 0.0834 \exp(+0.21 \text{ eV}/kT) \quad \text{for B} \tag{8.48a}$$

$$f_I = 0.0147 \exp(+0.40 \text{ eV}/kT) \quad \text{for P} \tag{8.48b}$$

$$f_I = 0.35 \quad \text{for As} \tag{8.48c}$$

$$f_I = 0.01 \quad \text{for Sb} \tag{8.48d}$$

It should be noted that these values are not definitive because of limited experimental data. However, these results clearly show that the fractional contribution of the interstitial-assisted mechanism to the total impurity diffusion is significant for P and B diffusion and is more significant for lower anneal temperatures. On the other hand, the vacancy-assisted mechanism must be dominant for Sb diffusion in Si.

8.8 Diffusion in Strained Silicon

The use of strained layers or heterostructures in Si CMOS has introduced new challenges in understanding and modeling dopant diffusion. Strain can alter the energy cost of many major steps involved in dopant diffusion: formation of native defects, displacement of a dopant atom to form a mobile dopant complex and probably clustering of these defects to form extended defects as well [47]. Strain-related band gap narrowing can also change the charged point-defect concentration [48]. However, there are many open questions about the effect of strain on dopant diffusion, such as the impact on the diffusion pathway and migration barrier.

Many studies have been done on the strain/stress effect on diffusion, theoretically, and experimentally [47–54]. Due to the complexity of such experiments, there are many inconsistent, even controversial, experimental results on the effect of stress on diffusion. In the mid-1990s, it was found that B diffusion is suppressed in strained SiGe [48,49]. Cowern et al. [5] proposed that the slower B diffusion is because of the biaxial compressive strain in the SiGe layer grown on the Si substrate. However, Kuo et al[51] found that even for relaxed SiGe, this diffusivity suppression still persists and that B diffusion exhibits weak strain dependence. Their results indicated that B diffusivity in Si reduces under tensile strain, which is opposite to the enhancement of B diffusivity in tensile-strained SiGe. Recently, Zangenberg et al. [52] reported results on strained SiGe and strained Si, which suggested that tensile strain increases the diffusion coefficient of B. One of the reasons for the large range of experimental data is the difficulty of decoupling stress effect and Ge chemical effect. Only the most recent data has clearly attempted to decouple the two factors. In addition, the presence of dilute levels of C in SiGe can dramatically reduce B TED because C getters Si interstitials. This has been exploited in the SiGeC base of heterojunction bipolar transistors.

Researchers have attempted to study B diffusion under stress in other systems. Zhao et al. [53] reported an enhancement of B diffusion under hydrostatic pressure and tried to predict biaxial strain effect on B diffusion by a theoretical approach with a strain-induced activation enthalpy, which relates to the activation volume found in hydrostatic pressure experiments. Actually, hydrostatic pressure is a tri-axial stress, which is quite different from the biaxial strain in strained Si/SiGe structure. In such a complicated system, it is difficult to make such comparisons.

Daw et al. [54] have presented a general treatment to obtain the dopant diffusivity tensors under strain. However, the basic parameters for this theoretical treatment cannot be easily obtained from experiments; they must be obtained from first-principles calculations. The recently proposed B diffusion pathway in Si [55,56] makes such a detailed first-principles study possible. Laudon et al. [57,58] studied B diffusion under TiN metal gate, which introduces compressive stress under the gate. They investigated the change of "creation volume", which is calculated by the length change, ΔL_α, between the defective cell and the perfect Si cell. However, they only did the calculation in detail for B diffusion under hydrostatic pressure. In addition, the relaxation of the defects is due not only to the change in volume of the cell, but also to the shape of the cell, which cannot be measured by "creation volume". Furthermore the process-induced uniaxial strain in Si, for example, under a nitride layer or TiN metal gate is not as uniform as in biaxially strained Si grown epitaxially on SiGe. It is a function of distance from the interface, and decays quickly down to the Si substrate. Biaxially strained Si grown on relaxed SiGe is a good candidate to study B diffusion under strain, experimentally, and theoretically. Although many studies of B diffusion in SiGe and strain effect on B diffusion have been done, the first-principles study of B diffusion in strained Si, which is an essential topic for understanding strained Si devices, is still very limited.

The effect of biaxial tensile strain in Si grown on relaxed SiGe on B diffusion was studied theoretically by Lin et al. [60]. All the calculations were performed with VASP, which is an ab initio quantum-mechanical molecular dynamics simulator based on DFT [59]. The potential used for this calculation was a generalized gradient approximation (GGA) ultra-soft pseudo-potential [61]. The simulations were performed on a uniform grid of k points equivalent to a $4\times4\times4$ Monkhorst and Pack grid in the diamond cubic cell. The energy cut-off was 208 eV. The optimized Si lattice constant for GGA in this system was 5.457 Å. A 64-atom super cell was used. Although most of previous studies of B diffusion were performed in a 64-atom system, a 216-atom unit cell was also studied in order to clarify the cell size effect on the results because strain probably magnifies the error for small cell sizes. Although even in the 216-atom system, B concentration is not dilute enough when compared with the real B concentration in MOSFETs, the interaction between two B in the two neighboring 216-atom systems is small enough for this study. In a 216-atom cell the energy differences for different defect configurations only shifts by a small value (10–20 meV in the strain range of this work), when compared with 64-atom cell.

The positively charged B–I pair is the lowest energy B interstitialcy, but B diffusion is arguably dominated by neutral pairs [56,62]. The change of the diffusion barrier with strain is small when

compared with the original differences between neutral and charged systems. Therefore, only the neutral system was considered. The results suggest that tensile biaxial strain lowers the activation enthalpy, and furthermore causes an anisotropy of B diffusion. On the contrary, compressive biaxial strain increases the activation enthalpy. The conclusion about tensile strain is consistent with experimental data. Although, there is no direct experimental evidence related with compressively strained Si, there are experiments on strained SiGe that can provide some guidance.

An enhancement and anisotropy of B diffusion in biaxial tensile strained Si was found. The diffusion barrier along the strain plane (channel) reduces, while the barrier in the vertical direction (depth) remains unchanged. This anisotropy comes from the orientation dependence of the saddle-point in the diffusion pathway. The formation enthalpy of B–I pair also decreases in strained Si. According to Lin et al's calculations [60], for strained Si on a $Si_{0.8}Ge_{0.2}$ buffer layer, which is widely used in strained MOSFET, an enhancement of B diffusivity along the channel by a factor ~ 4 and a factor ~ 2 in the vertical direction are expected for typical rapid thermal anneals (RTA).

8.9 Conclusions and Future Research

Modern theories of diffusion have recognized that to explain phenomena such as TED, it is important to explicitly take into account the non-equilibrium concentration of point defects such as vacancies and interstitials, both neutral and in various charged states [63–67]. While the concentration of the neutral species depends only on temperature, those of the charged states depend both on temperature and on doping. Existing simulators such as SUPREM4, in fact, are based on interactions with neutral and charged vacancies. More sophisticated simulators such as FLOOPS [32] take into account both vacancies and interstitials in various charged states. One writes the non-equilibrium diffusivity, D as

$$D = D_{\mathrm{V}}\left(\frac{C_{\mathrm{V}}}{C_{\mathrm{V}}^{*}}\right) + D_{\mathrm{i}}\left(\frac{C_{\mathrm{i}}}{C_{\mathrm{i}}^{*}}\right) + \text{other terms corresponding to charged defects,}$$

where C_{V} and C_{V}^{*} are the non-equilibrium and equilibrium concentrations of vacancies (similarly for the interstitials), and D_{V} and D_{i} are the diffusivities of dopant–vacancy (interstitial) pairs. It is then necessary to solve *coupled* partial differential equations for the concentrations of different types of point defects, taking into account the recombination of vacancy–interstitial pairs in the bulk and at the surface (especially for ultra-shallow junctions), the coalescence of the point defects into extended defects such as dislocation loops or {311} rod-like defects through so-called Ostwald ripening, and subsequent dissolution of these extended defects during annealing. The partial differential equations describing Fickian diffusion of the various dopant–defect pairs (charged and neutral) can be used to explain transient enhanced (or retarded) diffusion corresponding to a super-saturation (or deficit) of the various point defects. Clearly, one needs an accurate spatial and temporal description of these point defects in order to be able to model the dopant profiles under a wide range of processing conditions. While this basic philosophy is correct, we believe that the existing modeling approaches are deficient in that there are too many fitting parameters (the various values of D for the different dopant–defect pairs).

We, therefore believe, that there is a critical need for *ab initio* calculations of the dopant diffusion, using, for example, Hohenberg–Kohn–Sham DFT [68], and molecular dynamics (MD), based on the Car–Parrinello method [69]. The method is based on determining the formation enthalpies of different dopant–defect configurations such as dopant–interstitial pairs, and calculating the activation energies of diffusion of such pairs from first principles. The basic principle behind DFT is to describe a many-electron system in terms of its ground-state (electron) density [68]. This implies that every observable of a stationary quantum-mechanical system can be written as a functional of the ground-state density alone, and can be calculated using a variational principle from the one-particle density. In essence, this allows the many-body problem to be recast as a one-particle Schrodinger equation with a Hohenberg–Kohn–Sham self-consistent "effective" potential in the so-called local density approximation (LDA). One can

then combine these approaches with MD. Molecular dynamics is a very powerful tool, which only makes assumptions about the validity of classical, Newtonian mechanics and the Born–Oppenheimer (BO) approximation to describe ionic motion. The chief weakness with many MD simulations is in somewhat empirical choice of the interatomic potentials. We feel that DFT will provide a better approximation to these potentials. Once one has accurate values of these parameters, one can use them in the process simulators such as FLOOPS [32].

Acknowledgments

The author acknowledges the help of his former graduate students, K. Park, S. Batra, A. Sultan, T. Kirichenko, and L. Lin in this effort. His research in this area has been funded by SRC, SEMATECH and the Texas Advanced Technology/Research Program.

References

1. Plummer, J. D., M. Deal, and P. Griffin. *Silicon VLSI Technology*. Englewood Cliffs: Prentice Hall, 2000.
2. Tasch, A. and S. Banerjee. *Nuc. Inst. Methods B* 112 (1996): 177–183.
3. International Technology Roadmap for Semiconductors (published by STA), http://public.itrs.net/
4. Lanoo, M. and J. Bourgoin. *Point Defects in Semiconductors I: Theoretical Aspects*, Springer Series in Solid-State Sciences. 22. Berlin: Springer, 1981 2; Bourgoin, J., and J. Corbett. "Lattice Defects in Semiconductors." *Inst. Phys. Conf. Ser.* 23 (1975): 149.
5. Swalin, R. A. "Theoretical Calculations of the Enthalpies and Entropies of Diffusion and Vacancy Formation in Semiconductors." *J. Phys. Chem. Solids* 18 (1961): 290.
6. Seeger, A. and M. Swanson. In *Lattice Defects in Semiconductors*, edited by R. Hasiguti, 93. Tokyo: University of Tokyo, 1968.
7. Benneman, K. H. "New Method for Treating Lattice Point Defects in Covalent Crystals." *Phys. Rev.* A137 (1965): 1497.
8. Whan, R. E. "Investigations of Oxygen-Defect Interactions between 25 and 700K in Irradiated Germanium." *Phys. Rev.* A140 (1965): 690.
9. Fairfield, J. M. and B. J. Masters. "Self-Diffusion in Intrinsic and Extrinsic Silicon." *J. Appl. Phys.* 38 (1967): 3148.
10. Shockley, W. and J. Last. "Statistics of the Charge Distribution for a Localized Flaw in a Semiconductor." *Phys. Rev.* 107 (1957): 392.
11. Glyde, H. R. "Rate Processes in Solids." *Rev. Mod. Phys.* 39 (1967): 373.
12. Fahey, P. M., P. B. Griffin, and J. D. Plummer. "Point Defects and Dopant Diffusion in Silicon." *Rev. Mod. Phys.* 61 (1989): 289.
13. Dorner, P., W. Gust, B. Predel, U. Roll, A. Lodding, and H. Odelius. "Investigations by SIMS of the Bulk Impurity Diffusion of Ge in Si." *Phil. Mag.* 49 (1984): 557.
14. Lin, A. M., D. A. Antoniadis, and R. W. Dutton. "The Oxidation Rate Dependence of Oxidation-Enhanced Diffusion of Boron and Phosphorus in Silicon." *J. Electrochem. Soc.* 128 (1981): 1131.
15. TSUPREM4, is copyrighted software from Synopsis.
16. Miyake, M. "Oxidation-Enhanced Diffusion of Ion-Implanted Boron in Silicon in Extrinsic Conditions." *J. Appl. Phys.* 57 (1985): 1861.
17. Fair, R. B., M. L. Manda, and J. J. Wortman. "The Diffusion of Antimony in Heavily Doped and n- and p-type Silicon." *J. Mater. Res.* 1 (1986): 705.
18. Fair, R. B. and J. C. C. Tsai. "A Quantitative Model for the Diffusion of Phosphorus in Silicon and the Emitter Dip Effect." *J. Electrochem. Soc.* 124 (1977): 1107.
19. Fahey, P., R. W. Dutton, and S. M. Hu. "Supersaturation of Self-Interstitials and Undersaturation of Vacancies during Phosphorus Diffusion in Silicon." *Appl. Phys. Lett.* 44 (1984): 777.
20. Michel, A. E., W. Rausch, and P. A. Ronsheim. "Implantation Damage and the Anomalous Transient Diffusion of Ion-Implanted Boron." *Appl. Phys. Lett.* 51 (1987): 487.

21. Servidori, M., R. Angelucci, F. Cembali, P. Negrini, S. Solmi, P. Zaumseil, and U. Winter. "Retarded and Enhanced Dopant Diffusion in Silicon Related to Implantation-Induced Excess Vacancies and Interstitials." *J. Appl. Phys.* 61 (1987): 1834.
22. Antoniadis, D. A., A. M. Lin, and R. W. Dutton. "Oxidation Enhanced Diffusion of Boron and Phosphorus in Near-Intrinsic (100) Silicon." *Appl. Phys. Lett.* 33 (1978): 1030.
23. Taniguchi, K., K. Kurosawa, and M. Kashiwagi. "Oxidation-Enhanced Diffusion of Boron and Phosphorus in (100) Silicon." *J. Electrochem. Soc.* 127 (1980): 2243.
24. Mizuo, S. and H. Higuchi. "Retardation of Sb Diffusion in Si during Thermal Oxidation." *Jpn. J. Appl. Phys.* 20 (1980): 739.
25. Tan, T. Y. and U. Gosele. "Oxidation-Enhanced or Retarded Diffusion and the Growth or Shrinkage of Oxidation-Induced Stacking Faults in Silicon." *Appl. Phys. Lett.* 40 (1982): 616.
26. Lin, A. M., R. W. Dutton, D. A. Antoniadis, and W. A. Tiller. "The Growth of Oxidation Stacking Faults Point Defect Generation at Si–SiO_2 Interface during Thermal Oxidation of Silicon." *J. Electrochem. Soc.* 128 (1981): 1121.
27. Morehead, F. F. and R. F. Lever. "Enhanced 'Tail' Diffusion of Phosphorus and Boron in Silicon: Self-Interstitial Phenomena." *Appl. Phys. Lett.* 48, no. 2 (1986): 151.
28. Mulvaney, B. J. and W. B. Richardson. "Model for Defect–Impurity Pair Diffusion in Silicon." *Appl. Phys. Lett.* 51, no. 18 (1987): 1439.
29. Orlowski, M. "Advanced Diffusion Models for Submicron Technologies." *IEDM Tech. Dig.* (1988): 632.
30. Mulvaney, B. J. and W. B. Richardson. "The Effect of Concentration-Dependent Defect Recombination Reactions on Phosphorus Diffusion in Silicon." *J. Appl. Phys.* 67 (1990): 3197.
31. Giles, M. D. "Defect-Coupled Diffusion at High Concentrations." *IEEE Trans. Computer-Aided Design* 8, no. 5 (1989): 460.
32. Mulvaney, B. J., W. B. Richardson, and T. L. Crandle. "PEPPER-A Process Simulator for VLSI." *IEEE Trans. Computer-Aided Design* 8, no. 4 (1989): 336 (FLOOPS is available from Prof. M. Law, University of Florida).
33. Hu, S. M., P. Fahey, and R. W. Dutton. "On Models of Phosphorus Diffusion in Silicon." *J. Appl. Phys.* 54, no. 12 (1983): 6912.
34. Fair, R. B. "Low-Thermal-Budget Process Modeling with the PREDICT™ Computer Program." *IEEE Trans. Elec. Dev.* 35 (1988): 285.
35. Michel, A. E. "Rapid Annealing and the Anomalous Diffusion of Ion Implanted Boron into Silicon." *Appl. Phys. Lett.* 50, no. 7 (1987): 416.
36. Angelucci, R., P. Negrini, and S. Solmi. "Transient Enhanced Diffusion of Dopants in Silicon Induced by Implantation Damage." *Appl. Phys. Lett.* 49, no. 21 (1986): 1468.
37. Kim, Y. M., G. Q. Lo, and D. L. Kwong. "Anomalous Transient Diffusion of Boron Implanted into Preamorphized Si during Rapid Thermal Annealing." *Appl. Phys. Lett.* 55, no. 22 (1989): 2316.
38. Orlowski, M. "Impurity and Point Defect Redistribution in the Presence of Crystal Defects." *IEDM Tech. Dig.* (1990): 729.
39. Mathiot, D. and J. C. Pfister. "Dopant Diffusion in Silicon: A Consistent View Involving Non-Equilibrium Defects." *J. Appl. Phys.* 55 (1984): 3518.
40. Mathiot, D. and S. Martin. "Modeling of Dopant Diffusion in Silicon: An Effective Diffusivity Approach Including Point-Defect Couplings." *J. Appl. Phys.* 70 (1991): 3071.
41. Rorris, E., R. R. O'Brien, F. F. Morehead, R. F. Lever, J. P. Peng, and G. R. Srinivasan. "A New Approach to the Simulation of the Coupled Point Defects and Impurity Diffusion." *IEEE Trans. Computer-Aided Design* 9 (1990): 1113.
42. Morehead, F. F., E. Rorris, R. R. O'Brien, R. F. Lever, J. P. Peng, and G. R. Srinivasan. "Fast Simulation of Coupled Defect-Impurity Diffusion in FINDPRO: Comparison with SUPREM-IV and Other Programs." *J. Electrochem. Soc.* 138 (1990): 3789.
43. Orlowski, M. "Unified Model for Impurity Diffusion in Silicon." *Appl. Phys. Lett.* 53 (1988): 1323.

44. Baccus, B., T. Wada, N. Shigyo, M. Norishima, H. Nakajimi, K. Inou, T. Iinuma, and H. Iwai. "A Study of Non-Equilibrium Diffusion Modelling—Applications to Rapid Thermal Annealing and Advanced Bipolar Technologies." *IEEE Trans. Elec. Dev.* 39 (1992): 648.
45. Swaminathan, B., K. C. Saraswat, R. W. Dutton, and T. I. Kamins. "Diffusion of Arsenic in Polycrystalline Silicon." *Appl. Phys. Lett.* 40 (1982): 795.
46. Mei, L. and R. W. Dutton. "A Process Simulation Model for Multilayer Structures Involving Polycrystalline Silicon." *IEEE Trans. Elec. Dev.* 29 (1982): 1726.
47. Chaudhry, S. and M. Law. *J. Appl. Phys.* 82, no. 3 (1997): 1138.
48. Moriya, N., L. C. Feldman, H. S. Luftman, C. A. King, J. Bevk, and B. Freer. *Phys. Rev. Lett.* 71, no. 6 (1993): 883.
49. Kuo, P., J. L. Hoyt, J. F. Gibbons, J. E. Turner, R. D. Jacowitz, and T. I. Kamins. *Appl. Phys. Lett.* 62, no. 6 (1993): 612.
50. Cowern, N. E. B., P. C. Zalm, P. van der Sluis, D. J. Gravesteijn, and W. B. de Boer. *Phys. Rev. Lett.* 72, no. 16 (1994): 2585.
51. Kuo, P., J. L. Hoyt, J. F. Gibbons, J. E. Turner, and D. Lefforge. *Appl. Phys. Lett.* 66, no. 5 (1995): 580.
52. Zangenberg, N. R., J. Fage-Pedersen, J. L. Hansen, and A. N. Larsen. *J. Appl. Phys.* 94, no. 6 (2003): 3883.
53. Zhao, Y., M. J. Aziz, H.-J. Gossmann, S. Mitha, and D. Schiferl. *Appl. Phys. Lett.* 74, no. 1 (1999): 31.
54. Daw, M. S., W. Windl, N. N. Carlson, M. Laudon, and M. P. Masquelier. *Phys. Rev. B* 64 (2001): 045205.
55. Sadigh, B., T. J. Lenosky, S. K. Theiss, M.-J. Caturla, T. D. de la Rubia, and M. A. Foad. *Phys. Rev. Lett.* 83, no. 21 (1999): 4341.
56. Windl, W., M. M. Bunea, R. Stumpf, S. T. Dunham, and M. P. Masquelier. *Phys. Rev. Lett.* 83, no. 21 (1999): 4345.
57. Laudon, M., N. N. Carlson, M. P. Masquelier, M. S. Daw, and W. Windl. *Appl. Phys. Lett.* 78, no. 2 (2001): 201.
58. Windl, W., M. Laudon, N. N. Carlson, and M. S. Daw. *Comput. Sci. Eng.* 3, no. 4 (2001): 92.
59. Kresse, G. and J. Hafner. *Phys. Rev. B* 47 (1993): 558; Kresse, G., and J. Hafner. *Phys. Rev.* 49 (1994): 14251; Kresse, G., and J. Furthemuller. *Comput. Mater. Sci.* 6 (1996): 15; Kresse, G., and J. Furthemuller. *Phys. Rev. B* 54 (1996): 11169.
60. Lin, L., T. Kirichenko, S. K. Banerjee, and G. S. Hwang. "Boron Diffusion in Strained Si: A First-Principles Study." *J. Appl. Phys.* 96, no. 10 (2004): 5543–5547.
61. Perdew, J. P., K. Burke, and M. Ernzerhof. *Phys. Rev. Lett.* 77 (1996): 3865.
62. Jeong, J. W. and A. Oshiyama. *Phys. Rev. B* 64 (2001): 235204.
63. Kirichenko, T. A., S. K. Banerjee, and G. S. Hwang. "Interaction of Neutral Vacancies and Interstitials with the Si(001) Surface." *Phys. Rev. B* 70, no. 4 (2004): 45321.
64. Sultan, A., S. Banerjee, S. List, and M. Rodder. *Appl. Phys. Lett* 69, no. 15 (1996): 2228.
65. Sultan, A., M. Craig, K. Reddy, S. Banerjee, and L. Larson. *App. Phys. Lett.* 67, no. 9 (1995): 1223–1225.
66. Craig, M., A. Sultan, S. Banerjee, E. Ishida, and L. Larson. *11th International Conference on Ion Implant Technology*, 665–8.
67. Sultan, A., S. Banerjee, S. List, G. Pollack, and H. Hosack. *11th International Conference on Ion Implant Technology*, 25–8, 1996.
68. Gross, E. and S. Kurth. In *Relativistic & Electron Correlation Effects in Molecules & Solids*, edited by G. Malli, New York: Plenum, 1994.
69. Car, R. and M. Parrinello. *Phys. Rev. Lett.* 55, no. 22 (1985): 2471.

9

Oxidation and Gate Dielectrics

9.1 Introduction to Oxidation and Gate Dielectric Technology **9**-1
Furnaces and Rapid Thermal Processors • Oxidation and Annealing Ambient • Interface Control—Addition of Halogen Chemistry

9.2 Oxidation Theory **9**-5
Near-Atmospheric Processing in O_2 and H_2O • High/Low Pressure Effects

9.3 Oxidation Interactions **9**-12
Segregation at Oxide Interfaces • Crystal Orientation and Defect Effects • Silicon Oxynitride Films

9.4 Numerical Modeling of Oxidation **9**-13
Current Numerical Models • Limitations of Models

9.5 Metrology of Dielectric Films **9**-14
Film Thickness Measurement Techniques • Gate Oxide Integrity (GOI) and Measurement Techniques • Oxide Charge, Bulk Trap and Interface Trap Measurements • Issues with Ultra-Thin Film Measurements

9.6 Gate Dielectrics **9**-23
Plasma Nitrided Oxide • High-*k* Gate Dielectrics

9.7 Summary **9**-32
Future of SiON as a Gate Dielectric • Options for SiON Gate Dielectric Replacement

References **9**-33

C. Rinn Cleavelin
Luigi Colombo
Hiro Niimi
Sylvia Pas
Texas Instruments, Inc.

Eric M. Vogel
University of Texas at Dallas

9.1 Introduction to Oxidation and Gate Dielectric Technology

At the heart of the Semiconductor Industry and complimentary metal oxide semiconductor (CMOS) device technology has been the ability to produce and scale silicon dioxide (SiO_2) for silicon integrated circuit technology as it has progressed from submicron dimensions and recently to nanoscale dimensions, with gate dielectric thickness scaling in the sub-1.0 nm regime. As noted in the most recent Semiconductor Industry Associate (SIA) 2005 Edition of the International Technology Roadmap for Semiconductors (ITRS) [1], the scaling of the historical silicon dioxide or more recently silicon oxynitride (SiON) gate dielectrics into the sub-1.0 nm thickness regime leads to excessive gate leakage and requires new gate dielectric technology that has a higher dielectric constant. This allows a larger physical thickness or effective oxide thickness (EOT) to decrease the gate leakage due to tunneling through the gate dielectric.

Although several methods, such as wet and dry thermal oxidations, plasma anodization, vapor phase reaction, and wet anodization have been developed and implemented for the growth of silicon dioxide films, the dominant technology has been thermal oxidation due to its effectiveness for *gate oxides* in field effect devices and at the time of this writing, it will most likely continue to be needed for interface control and quality as dielectrics with higher dielectric constants are introduced.

In this chapter, we discuss and review the basic technology and theory of silicon oxidation; extending silicon-based dielectrics into the nanoscale regime by nitridation of the silicon dioxide; and transitioning of the silicon-based dielectric to a metal-based dielectric, such as hafnium (Hf) to significantly increase its dielectric constant. Although silicon dioxide serves functions other than gate oxide in integrated circuit fabrication, such as screening oxides, sidewall spacers, diffusion masks, surface passivation, interlevel dielectrics, and others; the continued use of silicon dioxide or a derivative of it for gate dielectric formation will continue to be a topic of research for the foreseeable future. The role, if any, that silicon oxidation or nitridation may play at the silicon channel or electrode interfaces is not totally clear at this time, but significant progress since the last edition of this handbook has been made.

Thus, this chapter will not only discuss the traditional silicon dioxide technology and its enhancement by nitridation, but it will also discuss in some detail the not too distant future of gate dielectrics incorporating high-k dielectrics and the challenges faced in implementing these films.

9.1.1 Furnaces and Rapid Thermal Processors

Although significant move furnacesment from batch to single wafer thermal tools has occurred over the last few years, batch thermal processors that can process many wafers at a time will continue to be used to reduce cycle-time and cost for processes which do not require the attributes associated with single wafer processors, such as rapid heating and cooling. In order to enhance the temperature ramping and uniformity in a conventional vertical furnace, furnace equipment manufacturers have introduced smaller batch vertical furnaces with smaller thermal masses that allow significantly faster heating and cooling, albeit, much more slowly than single wafer rapid thermal processors (RTP) systems. Shown in Figure 9.1a is an example of rapid thermal vertical oxidation furnace manufactured by Tokyo Electron Incorporated (TEL) called the TEL Formula with a line drawing (Figure 9.1b) showing the construction of the reaction chamber. Typical wafer loads for these furnaces are 25–100 wafers depending on the process requirements. To meet uniformity requirements, these furnaces utilize wafer rotation, multi-zone heater control, and gas distribution control in the vertical axis.

The typical furnace reactors consist of a high mass resistance-heated element typically operating in the 700°C–1100°C temperature range for oxidation processes. The TEL Formula employs a low-thermal mass heater to provide fast ramp up/down capability. This allows improved process cycle-time and lower cost-of-ownership (CoO) for processes that need to ramp over a significant temperature range. A typical vertical furnace has several key subsystems including (1) a quartz or SiC reaction chamber incorporating a quartz or SiC wafer tower, resistance-heated element with temperature control operating in the 700°C–1100°C temperature range for oxidation processes, and a quartz pedestal to support the wafer tower; (2) a gas and/or liquid injection system normally controlled by mass flow controllers (MFC) interfaced to the furnace computer; (3) a wafer handling system that includes mechanisms for cassette unloading and loading, cassette storage, wafer loading and unloading into the furnace tower, and a mechanism to insert and remove the wafer tower into the reaction chamber; (4) controlled exhaust or vacuum pumping system to remove reacted and unreacted gases from the reaction chamber and reduced pressure operation if required; (5) tower (wafer) rotation mechanism if needed for improved azimuthal uniformity; and (6) a laminar flow area that encloses the cassette, wafer, and tower handling mechanisms to minimize any particulate contamination to the wafers during wafer handling and storage. In addition to these subsystems, some of today's processes are utilizing either vacuum or nitrogen load–lock systems for improved interface control.

A typical oxidation process consists of a pre-oxidation cleaning (see Chapter 4) step followed by an immediate loading of the cleaned wafers into the furnace's cassette storage area. As mentioned above,

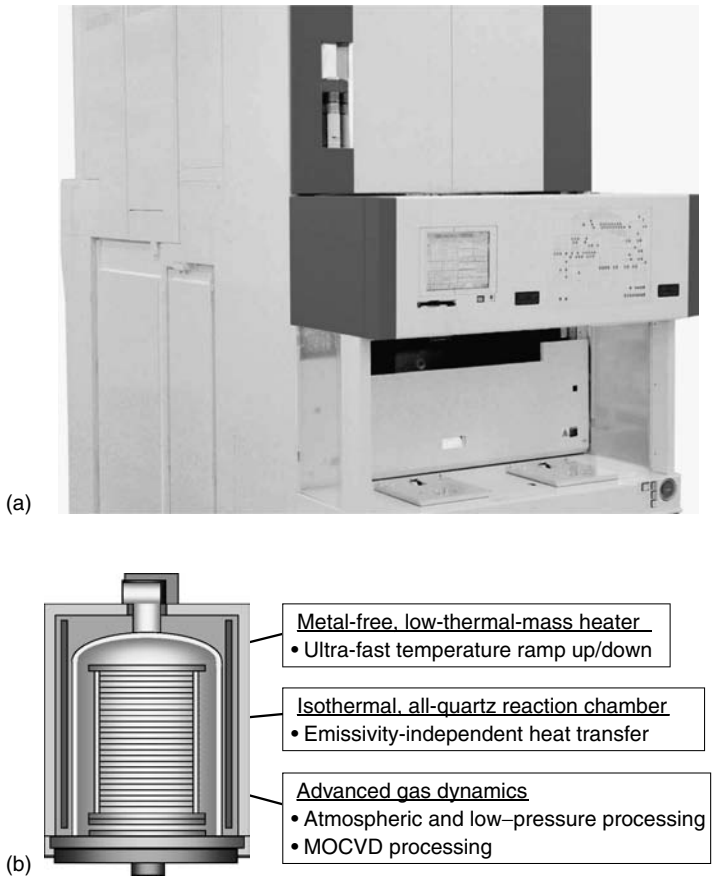


FIGURE 9.1 (a) TELFormula™ multi-wafer thermal reactor. (b) Line drawing showing the construction of the reaction chamber. (Courtesy of Tokyo Electron, Inc.)

in some furnaces this may be through a nitrogen (N_2) or vacuum load lock to prevent oxygen (O_2) or moisture from the ambient from growing excessive native oxide during wafer loading and pushing of the wafer tower into the heated environment of the furnace reaction chamber. Other techniques, such as fast ramp furnace elements are also used to prevent native oxidation. This allows the wafers to be pushed into the reaction chamber at low enough ($<450^{\circ}C$) temperatures so that little or no native oxide is grown during the insertion into the tower. As a result, cycle-time is minimized by ramping the temperature from the lower loading- to process temperature. After the wafer tower is loaded into the reaction chamber, usually in a non-oxidizing ambient, the furnace temperature is ramped to process temperature and an oxidizing ambient is introduced into the reaction chamber through quartz injectors. The typical oxidizing ambients are oxygen (O_2) or a mixture of H_2 and O_2 reacted in an external pyrogenic torch assembly to produce water vapor (steam). Since the reaction of the O_2 or steam is diffusion limited, a uniform oxide growth for oxide layers thicker than 100 Å is easily achieved after the first few monolayers of oxide. For thin oxides, much better care must be taken to ensure the initial oxide growth is uniform, since it will greatly influence the final film uniformity as well as the film's electrical reliability characteristics. Some of the key influences on film uniformity are surface preparation, storage ambient control, gas flow dynamics in the reaction chamber, temperature uniformity across the wafers and down the reaction chamber, temperature ramp control, gas purity, wafer spacing (pitch), and post-oxidation ambient and temperature treatments. For ultra-thin oxidation (1–4 nm), vacuum or

nitrogen load-locks, low temperature loading, in-situ surface treatments, and/or wafer rotation are required to achieve uniformity and oxide quality control.

9.1.1.1 Rapid Thermal Oxidation

With the introduction of single wafer rapid thermal oxidation (RTO) systems as shown in Figure 9.2a, in-situ surface preparation, wafer rotation, ambient control with toxic gas capability, and process integration become much more tenable. The capability of controlling and having full flexibility of the H_2:O_2 mixture ratio from an oxygen-rich to a hydrogen-rich ambient allows oxidation parameters to be varied from a rapid oxidation growth to a selective oxidation regime. Rapid thermal oxidation becomes capable of competing economically with the more conventional vertical hot-wall batch furnaces with rapid oxidation rates (O_2-rich) in excess of 100 nm/min. In the case of H_2-rich environments, single wafer RTO configuration provides this capability safely, since the reaction chamber volume is small and the process can be performed at reduced pressure limiting the reactive gases present, whereas, in a conventional batch furnace the volume is large and the risk of using H_2-rich ratios is a significant safety deterrent. Unlike a furnace with a pyrogenic torch, no hot quartz is used to generate the steam ambient thus eliminating quartz devitrification. In Figure 9.2b, the H_2O in-situ steam generation (ISSG™) is generated directly in the reaction chamber and primarily at the wafer surface by the introduction of electronic-grade H_2 and O_2. This configuration also reduces the risk of metal contamination since no metal catalyst is used in the steam generation.

Figure 9.2b is a cross-sectional view of the reaction chamber depicting the necessary components of a lamp heated RTO system (courtesy of Applied Materials, Inc.). In the case of this reactor, the wafer is inserted into the reaction chamber by a robot wafer handler onto a rotating SiC support ring. The wafer is rapidly heated at rates of 50°C–100°C/s to process temperature 800°C–1050°C. Process gases are introduced at reduced pressure in an axial flow pattern where they react to form water vapor, hydroxyl ions, oxygen ions, and other species depending on the H_2:O_2 ratio and whether N_2, NH_3, HCl, or other reactant gases are introduced. At the conclusion of the oxidation and annealing steps, the wafer is cooled rapidly to a temperature, where it can be safely removed by the wafer handler from the reaction chamber to an additional cool-down chamber that can be used to enhance system throughput. Cool-down can be enhanced further by providing a wafer backside cooling gas, such as helium or hydrogen. This capability is especially important if the oxidation step has a limited thermal budget. In the system shown in Figure 9.2a and Figure 9.2b measurement and uniformity control of the wafer temperature is accomplished using a linear array of fiber optic pyrometer probes with active feedback. In addition, the SiC support ring is

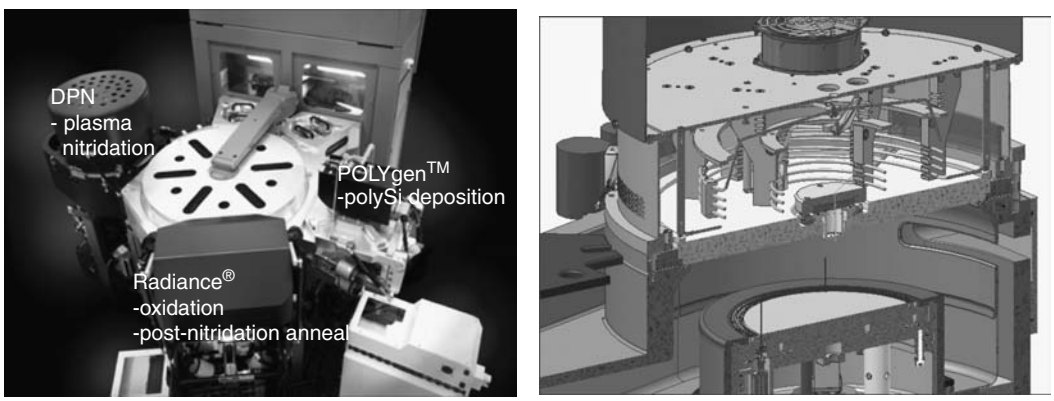


FIGURE 9.2 (a) A 300-mm Applied Centura® DPN gate stack. (b) A 300-mm DPN Chamber cut-away. (Courtesy of Applied Materials, Inc.)

necessary to prevent a large radial temperature gradient by supplying heat to the wafer's edge during temperature ramping, since radiative cooling is the largest at the wafer's edge. For toxic or corrosive gas operation, materials of construction and gas handling must be given careful design consideration.

9.1.2 Oxidation and Annealing Ambient

The largest impact of the annealing of silicon dioxides is for reduction of charge trapping and interface trap density whether this oxide is a thermal gate oxide or a deposited oxide [2]. Oxide annealing is normally performed immediately after the oxidation process and after the metal or polysilicon gate has been formed. In the latter case, this anneal is referred to as a post-metal anneal (PMA) or sinter [3]. In the case of post-oxidation anneal (POA), the temperature, pressure, and ambient can have an impact on V_{bd}, Q_{bd}, mobility, D_{it}, and oxide uniformity [4–8]. The POA ambient is typically N_2 or Ar. The PMA or sinter ambient is typically a mixture of N_2 and H_2 for hydrogen incorporation at the silicon–silicon dioxide ($Si–SiO_2$) interface to passivate dangling bonds [9]. The incorporation of nitrogen into the silicon dioxide using N_2O annealing continues to be under investigation as a barrier against boron diffusion [10]. Charge trapping and interface trap density are found to be reduced and charge-to-breakdown improved in this case as well. It has been reported that ΔD_{it} due to stress induced interface state generation can be reduced using a high temperature post-gate oxidation anneal in Ar [11]. Ajuria et al. show that POA significantly improved charge-to-breakdown and interface hardness at very low backend temperatures, but that there was little improvement at high backend temperatures [12]. They propose that the elimination of POA may have no adverse effects on gate oxides because thermal processing post-gate oxidation often includes high temperature processing, which could be substituted for POA in some cases.

Pre-oxidation annealing has been found to reduce defect density and improve mobility [13] as well as gate oxide film quality [14]. In addition, POA in N_2O was found to improve oxide reliability with low interface trap density, improved high field mobility, low charge trapping, and increased resistance to hot-carrier induced interface state generation. Annealing in various ambients was also investigated. Beck et al. have studied the effects of micro-contaminants in the annealing ambients on oxide quality [15].

9.1.3 Interface Control—Addition of Halogen Chemistry

Halogens, typically chlorine (Cl), have been used in silicon oxidation for several reasons. The addition of Cl is known to enhance oxidation rates [3,4] as well as increase resistance to gamma ray radiation damage [3]. Cl also reduces effects of metallic contamination at the $Si–SiO_2$ interface by tying Cl to the metal ions [10]. For example, Cl is recommended for oxidation of trenches for shallow trench isolation (STI) because of metallic contamination remaining from the preceding chemical mechanical polishing (CMP) process. Cl-based oxidation can enhance the oxidation rate depending on the dopant type and concentration and, therefore, cause a shift in threshold voltage (Vt). The concentration of Cl in the process gases is an additional factor in oxidation rates and Vt shift. Acceptable manufacturing sources of Cl include 1,1,1-Trichloroethane (TCA), HCL, 1,2-Dichloroethylene (*t*-DCE) [16], and trans-Dichloroethylene (trans-LC). In general, use of halogens in semiconductor manufacturing is discouraged because of fab safety, equipment corrosion, and environmental impact. TCA has been identified as an ozone depleting compound [16].

9.2 Oxidation Theory

9.2.1 Near-Atmospheric Processing in O_2 and H_2O

This section describes the mechanisms and kinetics associated with thermal oxidation of silicon at near atmospheric pressures. When a silicon surface is exposed to an oxidizing ambient, usually at atmospheric pressure and high temperatures, SiO_2 is formed. The chemical reaction, which describes "dry oxidation"

of silicon in pure oxygen, is

$$\mathrm{Si(solid)} + \mathrm{O_2} \leftrightarrow \mathrm{SiO_2(solid)} \tag{9.1}$$

The reaction describing "wet oxidation" of silicon in water vapor is

$$\mathrm{Si(solid)} + 2\mathrm{H_2O} \leftrightarrow \mathrm{SiO_2(solid)} + 2\mathrm{H_2} \tag{9.2}$$

In this process, silicon covalently bonds to oxygen through the sharing of valence electrons. The amount of available oxidant molecules, the amount of available silicon, and the rate of reaction between these reactants determine the rate of oxide growth. Since silicon is consumed during the oxidation, the final oxide layer can be calculated from the density and molecular weight of silicon and oxide to be approximately 54% above the original silicon surface and 46% below the original surface. It has been demonstrated that oxidation proceeds by diffusion of the oxidant through the oxide and the interaction of the oxidant at the Si–SiO_2 interface [17–19]. However, the nature of the interfacial reaction is still under debate.

9.2.1.1 Deal–Grove Silicon Oxidation Model

The Deal–Grove model describes thermal oxidation of silicon for oxide thicknesses ranging from 30 to 2000 nm, oxidant partial pressures between 0.1 and 1.0 atm, temperatures from 700 to 1300°C, under both pure oxygen and water vapor [20]. A cross-section of the Si–SiO_2 layer illustrating the basic Deal–Grove model for thermal oxidation is shown in Figure 9.3. For the occurrence of oxidation, the oxidant must: (1) travel from the gas phase to the gas-oxide interface with flux, (2) move across the SiO_2 film toward the silicon with flux F_2, and (3) react with silicon at the Si–SiO_2 interface with flux F_3. This model assumes that the oxidation process is in steady state, so that each of the fluxes must be equal. ($F = F_1 = F_2 = F_3$).

The gas-phase oxidant flux, F_1 is given as

$$F_1 = h(C^* - C_0) \tag{9.3}$$

where h is the gas phase transport coefficient, C_0 is the oxidant concentration in the oxide at the outer surface, and C^* is the equilibrium oxidant concentration in the oxide. Henry's law is used to

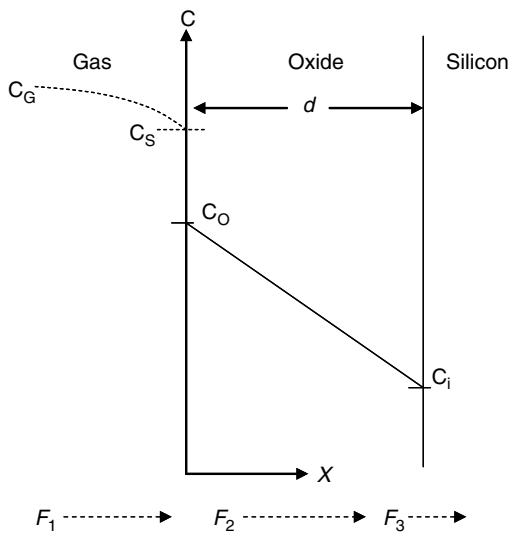


FIGURE 9.3 Model for the oxidation of silicon. (From Deal, B. E., *J. Electrochem. Soc.*, 125, 576, 1978.)

relate the partial pressure of the oxidant in the gas, p, with the equilibrium oxidant concentration in the oxide,

$$C^* = kp, \tag{9.4}$$

where k is Henry's law constant. Henry's law only holds when the oxidant does not associate or dissociate at the outer surface, which implies that the oxidant species is molecular. The observed linear dependence of oxidation rate on pressure supports this assumption.

Fick's law is used to describe the flux of oxidant across the oxide layer as,

$$F_2 = \frac{D_{\text{eff}}(C_0 - C_i)}{x_0}, \tag{9.5}$$

where D_{eff} is the effective diffusion constant, C_i is the oxidant concentration at the Si–SiO_2 interface, and x_0 is the oxide thickness.

The flux corresponding to the reaction of the oxidant at the Si–SiO_2 interface is given as,

$$F_3 = k_s C_i, \tag{9.6}$$

where k_s is the chemical reaction rate constant.

By solving the above equations for the flux, F, the oxide growth rate can be determined from

$$\frac{dx_0}{dt} = \frac{F}{N_1}, \tag{9.7}$$

where N_1 is the number of oxidant molecules incorporated into a unit volume of oxide layer and N_1 is 2.2×10^{22} SiO_2 molecules/cm^3 for dry oxygen and 4.4×10^{22} molecules/cm^3 for H_2O. Solving this differential equation and assuming that an initial thickness of oxide (x_i) may be present ($x_0 = x_i$ at $t = 0$), results in the following relationship for oxide thickness as a function of time,

$$x_0 = \frac{A}{2}\left[\left(1 + \frac{t+\tau}{A^2/4B}\right)^{1/2} - 1\right], \tag{9.8}$$

where,

$$A = 2D\left[\frac{1}{k_s} + \frac{1}{h}\right](\text{cm}) \tag{9.9}$$

$$B = \frac{2DC^*}{N_1}(\text{cm}^2/\text{s}) \tag{9.10}$$

$$\tau = \frac{x_i^2 + Ax_i}{B}(\text{s}) \tag{9.11}$$

τ represents an equivalent shift in time due to the presence of an initial oxide thickness. For long oxidation times ($t \gg \tau$, $t \gg A^2/4B$), Equation 9.8 reduces to the parabolic law

$$x_0^2 = Bt \tag{9.12}$$

where B is the parabolic rate constant. For short oxidation times ($t + \tau \ll A^2/4B$), Equation 9.8 reduces to the linear law

$$x_0 = \frac{B}{A}(t + \tau) \tag{9.13}$$

where B/A is the linear rate constant. For thick films (long oxidation times), the reaction is limited by the

TABLE 9.1 Rate Constants for Oxidation of Silicon in Wet Oxygen

Oxidation Temperature (°C)	Parabolic Rate Constant		Linear Rate Constant	
	A (μm)	B (μm^2/h)	B/A (μm/h)	(h)
1200	0.05	0.720	14.40	0
1100	0.11	0.510	4.64	0
1000	0.226	0.287	1.27	0
920	0.50	0.203	0.406	0

Source: From Ahn, S. T. et al., *J. Appl. Phys.*, 65, 1989.

oxygen diffusion step resulting in the parabolic growth vs. time dependence, whereas for thinner films (short oxidation times) the reaction is limited by the kinetics of the oxygen reaction with silicon resulting in the linear growth vs. time dependence.

Deal and Grove compared the model with experimental data taken on ⟨111⟩ lightly boron-doped silicon wafers. The results obtained are valid over the range of given conditions. The constants in Equation 9.9 through Equation 9.11 were determined graphically through manipulation of Equation 9.8. For wet oxidation in water vapor it was found that $x_i = 0$ for all temperatures. For dry oxidation it was found that extrapolation of x_0 to $t = 0$ resulted in an intercept of approximately 23 nm independent of temperature. Therefore, a value of $x_i = 23$ nm must be used in τ. This enhanced dry oxidation rate above the purely linear relationship for short oxidation times will be discussed further in this section. Values for A, B/A, and τ as a function of temperature are given in Table 9.1 and Table 9.2 and in Figure 9.4 and Figure 9.5. The results indicate that both the parabolic rate constant, B, and linear rate constant, B/A, increase exponentially with temperature. The temperature dependence of B can be traced to that of the effective diffusion coefficient, D_{eff} and that of B/A has been related to the surface reaction rate constant, k_s. The pressure dependence of the rate constants has also been determined [20,21]. The results indicate that A is independent of pressure, whereas, B is linearly dependent on pressure. This linear dependence of B on pressure can be attributed to the linear dependence of C^* on pressure in Equation 9.4. The crystal orientation dependence of the parabolic and linear rate constants indicate that the B/A for ⟨111⟩ silicon is an average of 1.68 times that of B/A for ⟨100⟩ silicon [20,22].

9.2.1.2 Thin Silicon Oxidation Models

The above Deal–Grove model for oxidation has provided good agreement over the given range of experimental conditions with the exception of dry oxidation of films with thickness less than approximately 30 nm. For thicknesses less than 30 nm the oxidation rate was observed to be faster than that predicted by the Deal–Grove model. As reviewed in Ref. [23], researchers have suggested many effects to explain the growth rate enhancement: electrochemical effects such as field-enhanced oxidation, structural effects such as microchannels, stress effects modifying the oxidant diffusivity, and changes in the oxygen solubility in the oxide. However, models which have provided the best qualitative and

TABLE 9.2 Rate Constants for Oxidation of Silicon in Dry Oxygen

Oxidation Temperature (°C)	Parabolic Rate Constant		Linear Rate Constant	
	A (μm)	B (μm^2/h)	B/A (μm/h)	(h)
1200	0.040	0.045	1.12	0.027
1100	0.090	0.027	0.30	0.076
1000	0.165	0.0117	0.071	0.37
920	0.235	0.0049	0.0208	1.40
800	0.370	0.0011	0.0030	9.0
700	—	—	0.00026	81.0

Source: From Ahn, S. T. et al., *J. Appl. Phys.*, 65, 1989.

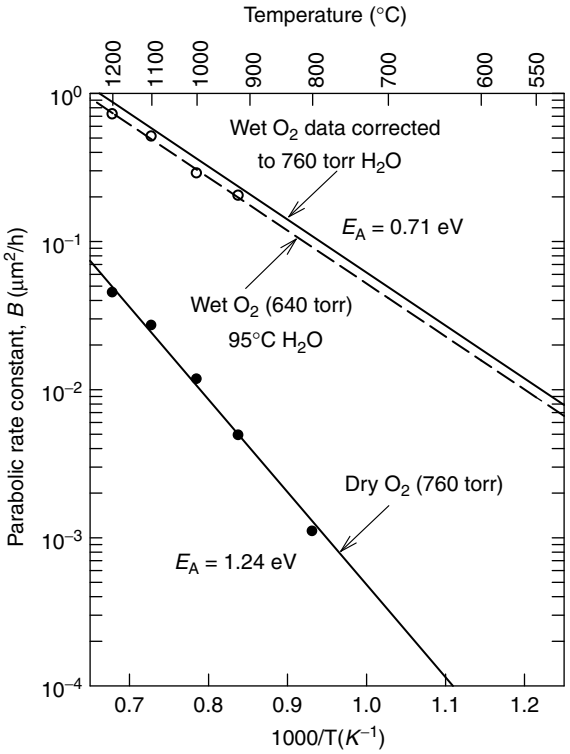


FIGURE 9.4 The effect of temperature on the parabolic rate constant, B for ⟨111⟩ silicon. (From Deal, B. E., *J. Electrochem. Soc.*, 125, 576, 1978.)

quantitative fits to experimental results are those based on parallel oxidation mechanisms, such as silicon interstitial injected into the oxide [24,25], oxygen vacancies [24], diffusion of atomic oxygen [25,26], surface oxygen exchange [24–26], and the effects of a finite non-stoichiometric transition region between amorphous SiO_2 and Si [25–27]. The following provides two such analytical models that have provided good fits to the data.

The research reported in Refs. [23,26,28] provided an analytical model based on parallel oxidation mechanisms to fit the experimental data. The following analytical model provided a good fit to experimental data,

$$x_0 = \left\{ \left(\frac{A}{2}\right)^2 + Bt + M_1\left[1-\exp\left(\frac{-t}{\tau_1}\right)\right] + M_2\left[1-\exp\left(\frac{-t}{\tau_2}\right)\right] + M_0 \right\}^{1/2} - \frac{A}{2} \tag{9.14}$$

where,

$$M_0 = \left(x_i^2 + Ax_i\right) \tag{9.15}$$

$$M_1 = K_1\tau_1 \tag{9.16}$$

$$M_2 = K_2\tau_2 \tag{9.17}$$

$$K_1 = K_1^0 \exp\left(\frac{-\Delta E_{K_1}}{kT}\right) \tag{9.18}$$

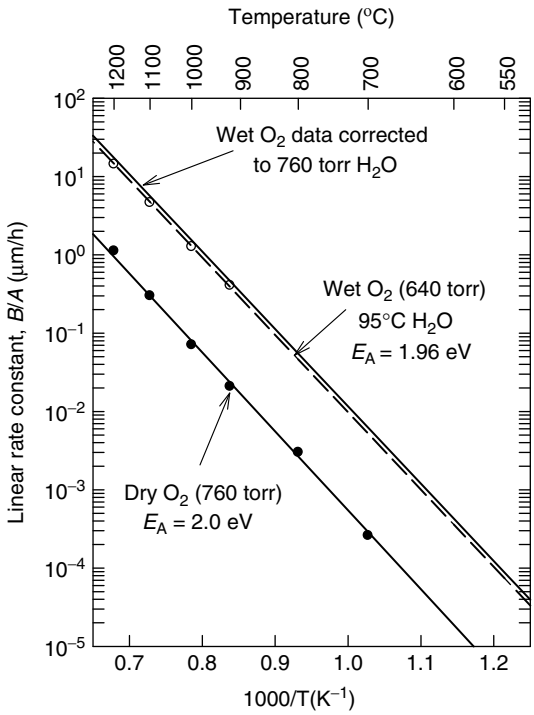


FIGURE 9.5 The effect of temperature on the linear rate constant, B/A for ⟨111⟩ silicon. (From Deal, B. E., *J. Electrochem. Soc.*, 125, 576, 1978.)

$$K_2 = K_2^0 \exp\left(\frac{-\Delta E_{K_2}}{kT}\right) \tag{9.19}$$

$$\tau_1 = \tau_1^0 \exp\left(\frac{-\Delta E_{\tau_1}}{kT}\right) \tag{9.20}$$

$$\tau_2 = \tau_2^0 \exp\left(\frac{-\Delta E_{\tau_2}}{kT}\right) \tag{9.21}$$

The values for the pre-exponential constants, time constants, and activation energies are given in Table 9.3. The initial thickness, x_i, is the native oxide thickness.

The authors in Ref. [29] also provided an analytical model based on parallel oxidation mechanisms to fit the experimental data that has slightly different form than Equation 9.14. The following thickness vs. time relationship was determined

$$\left(x_0^2 - x_i^2\right) + C(x_0 - x_i) - G \ln\left(\frac{2Ex_0 + F}{2Ex_i + F}\right) = Et \tag{9.22}$$

where,

$$C = \frac{A_1B_1 + A_2B_2}{B_1 + B_2} \tag{9.23}$$

$$E = B_1 + B_2 \tag{9.24}$$

TABLE 9.3 Arrhenius-Expression Parameters of the Time Pre-Exponential Constants and Time Constants of Equation 9.14 for Silicon Oxidized in Dry Oxygen in the 800°C–1000°C Range

Orientation	⟨100⟩	⟨111⟩	⟨110⟩
$Å^2$/min	2.49×10^{13}	2.70×10^{11}	4.07×10^{10}
eV	2.18	1.74	1.54
$Å^2$/min	3.72×10^{13}	1.33×10^{11}	1.20×10^{10}
eV	2.28	1.76	1.56
Min	4.14×10^{-6}	1.72×10^{-6}	5.38×10^{-9}
eV	1.38	1.45	2.02
Min	2.71×10^{-7}	1.56×10^{-7}	1.63×10^{-8}
eV	1.88	1.90	2.12

Source: From Wagner, C., *J. Appl. Phys.*, 29, 1295, 1958.

$$F = A_1B_2 + A_2B_1 \tag{9.25}$$

$$G = \frac{B_1B_2(A_1 - A_2)^2}{2(B_1 + B_2)} \tag{9.26}$$

The parameters were obtained from experimental data and are summarized in the following,

$$B_1(100) = B_1(111) = 6.5\times10^{11}\exp\left(-\frac{2.2\text{ eV}}{kT}\right)Å^2/\text{min} \tag{9.27}$$

$$A_1(100) = A_1(111) = 0 \tag{9.28}$$

$$B_2(100) = B_2(111) = 2.6\times10^{10}\exp\left(-\frac{1.6\text{ eV}}{kT}\right)Å^2/\text{min} \tag{9.29}$$

$$\frac{B_2}{A_2}(100) = 2.6\times10^{8}\exp\left(-\frac{1.9\text{ eV}}{kT}\right)Å/\text{min} \tag{9.30}$$

$$\frac{B_2}{A_2}(111) = 8.3\times10^{8}\exp\left(-\frac{1.9\text{ eV}}{kT}\right)Å/\text{min} \tag{9.31}$$

9.2.2 High/Low Pressure Effects

Thermal oxidation of silicon at pressures greater than 1 atm has been used to accelerate the oxide growth at lower temperatures, thereby minimizing dopant redistribution in the silicon. Furthermore, it has been shown that oxidation induced stacking faults are reduced when processing at higher pressures [30]. High-pressure oxidation has found applications mainly in the areas of field oxide growth and selective oxidation due to the thick oxides needed [30]. As shown in Section 9.2.1.1, the pressure dependence of both B and B/A for pressures of 0.1–1.0 atm was found to be linear. The researchers in Ref. [31] have shown that for oxidation of silicon in steam for pressures of 5–20 atm and temperatures of 800°C–1000°C, that both B and B/A are linearly proportional to pressure. However, for very large pressures (>25 atm), Ligenza has shown that the rates become nonlinear due to the volatility of the silicon oxide in steam [17]. Above a certain pressure for a given temperature, the film will no longer grow and an etching of the silicon occurs. The minimum steam pressure for this to occur is 500, 400, 200, 150 atm for temperatures of 500, 650, 750, and 850°C, respectively [17]. For dry oxidation, Lie et al. have shown that for pressures of

5–20 atm and temperatures of 800°C–1000°C that B is linearly proportional to pressure, whereas B/A is proportional to oxygen pressure to the power of 0.7 [32].

For very low oxygen partial pressures or temperatures near the melting point of silicon, the reaction given [33–35] in Equation 9.1 can proceed in a reduction mode whereby [36] SiO_2 can be etched or reduced by silicon to form SiO, which diffuses away as a gas [37,38]. Furthermore, POA at high temperatures and low oxidant partial pressures has also resulted in the etching or reduction of the oxide [33,34,39–41]. This reaction can cause many effects including vacancy formation in silicon [33], defect formation in the oxide [39,40,53], void growth [34,38,40], surface roughness [40], and reduction of oxide thickness [41].

9.3 Oxidation Interactions

9.3.1 Segregation at Oxide Interfaces

The presence and quantity of dopants at the Si surface can change oxidation kinetics. Oxidation rates typically increase with heavier doping [42] and the oxidation process itself changes the behavior of dopants [42,43] and their distribution in the silicon or polysilicon layer [44–46]. At low temperatures, where dopant diffusion in silicon and oxidation rates are very slow, high-pressure oxidation causes dopant out-diffusion into the growing oxide [47].

9.3.2 Crystal Orientation and Defect Effects

Crystal orientation changes the oxidation rate because of the number of open lattice sites at the Si–SiO_2 interface. Typically, ⟨111⟩ and ⟨110⟩ oriented single crystal silicon have increased oxidation rates over ⟨100⟩ orientation. The larger number of open sites also increases Si surface roughness and can easily cause Si plane defects to propagate through the growing oxide. More open lattice sites also translate to a higher number of dangling bonds at the Si–SiO_2 interface, which cause electrical defects. One defect that is critical in gate oxidation is crystal originated particles (COPs). These defects occur on the single crystal Si wafer and act as current leakage path in a gate oxide [48–50]. Crystal growing or polishing techniques directly influence COP formation.

9.3.3 Silicon Oxynitride Films

Silicon oxynitride films have been implemented in recent years due to some of their properties including higher barrier properties to impurity penetration, such as Boron, hot-carrier resistance, radiation damage resistance, and improved high field electron channel mobility and dielectric constant modification [51–58]. These films were mainly formed through the thermal nitridation of silicon oxide or the oxidation of silicon in nitrogen containing ambient, such as NH_3, N_2O, and NO. In recent years, industry has developed and implemented into production a plasma nitridation technique described below that provides a preferred nitrogen profile through the gate dielectric. A review of these main processes is discussed below. It should be mentioned that many of the experimental results associated with oxynitrides are equipment-specific. Furthermore, wide ranges of processing conditions have been used to form and modify silicon oxynitride. Unless otherwise stated, the following review represents results that are relatively generic with detailed experimental conditions. However, the details of results will vary depending on the experimental conditions.

9.3.3.1 Ammonia-Based Oxynitrides

Annealing of silicon dioxide in NH_3 usually results in a peak of nitrogen at the silicon–silicon oxynitride (Si–SiON) interface, a small amount of nitrogen in the bulk and a nitrogen rich surface region [59–62]. The nitrogen at the surface has been attributed to the exchange of O and N atoms during nitridation [61]. Nitridation proceeds through the diffusion of a nitrogenous species towards the Si–SiON interface [59, 63]. The nitrogen peak at the interface has been reported to be consistent with nitrogen atoms relieving interfacial strain and the replacing of weak Si–O bonds found near the interface. The amount of nitrogen in these three regions is strongly dependent on experimental conditions. Higher pressures, higher

temperatures, and longer times generally result in an increase in the nitrogen concentration especially at the Si–SiON interface. Nitrogen concentrations from 1 atm% upwards to approximately 40 atm% have been reported [55,59–63].

The formation of oxynitrides using NH_3 also introduces hydrogen into the film, which has been associated with dielectric reliability problems [54,55]. To remove this incorporated hydrogen, post-nitridation anneals in various ambients have been employed. Post-nitridation anneals in O_2 (reoxidation) have traditionally been the most popular and modify the film in several ways [59]. With increasing reoxidation temperature the hydrogen concentration decreases more rapidly. A similar rate of decrease with temperature was also observed for N_2 anneals suggesting that the reduction of hydrogen is mainly due to a diffusion mechanism rather than chemical reaction. It was observed that hydrogen removal is slower for higher nitrogen concentrations suggesting that nitrogen acts as a hydrogen diffusion barrier. With increased reoxidation time, the Si–SiO_2 interface oxidizes causing a movement of the starting interface position into the substrate and an increase in film thickness. The rate of thickness increase is lower for lower temperatures. Higher nitrogen concentrations typically slow this oxidation rate suggesting that nitrided oxides are a diffusion barrier to oxygen. The peak of nitrogen concentration moves with the oxidation keeping approximately the same distance away from the formed Si–SiO_2 interface. Reoxidation also decreases the bulk nitrogen concentration, while only slowly decreasing the interfacial nitrogen concentration. N_2 anneals are not as effective as O_2 anneals in removing bulk nitrogen suggesting that nitrogen removal is mainly due to an oxidation mechanism. The research by Baumvol et al. [61] and Carr et al. [64] suggests that nitrogen removal is due to an exchange mechanism of nitrogen in the film with oxygen introduced during reoxidation.

9.3.3.2 Nitrous and Nitric Oxide-Based Oxynitrides

The formation of silicon oxynitride using N_2O and/or NO results in a build-up of interfacial nitrogen and bulk nitrogen throughout the film. The nitrogen concentration and its distribution is very sensitive to process parameters, such as temperature, time, pressure, processing variations (e.g., annealing of a pre-oxide or oxynitridation of silicon), and tool considerations (furnace or RTP, associated gas-phase kinetics and thermodynamics, etc.) [51,65–71]. Furthermore, RTP have shown large compositional and thickness nonuniformities across the wafer for N_2O processing [68,72,73].

The interfacial nitrogen concentration found in N_2O nitrided oxides is typically less than NH_3 and NO nitrided oxides. This has been explained by the fact that NO is the species responsible for nitridation in both the NO and N_2O ambients [51,74]. It has also been observed that N_2O and NO annealing results in the removal of bulk nitrogen, which has been attributed to the presence of monoatomic oxygen in the annealing ambient [64]. Oxide growth kinetics associated with NO and N_2O have shown an initial stage of fast growth followed by a reduction in oxidation rate, which has been attributed to the nitrogen-rich interfacial layer acting to block the oxidant diffusion to the interface [65,67,69–71,75,76]. Several studies have modeled the growth kinetics of thermal oxynitridation of silicon in N_2O by modifying the Deal–Grove theory [75,76].

9.4 Numerical Modeling of Oxidation

9.4.1 Current Numerical Models

Numerical models for oxidation of silicon cover a range of techniques. A nonlinear visco-elastic model incorporated into the Deal–Grove kinetics model uses the implicit backward Euler method to show the time dependence and the finite element method to solve the oxidant concentration and oxide shape changes [77]. Nonlinear functions of stress are solved using iterations are typical. Finite element analysis in the program (NOVEL) is incorporated into the finite difference method used for process simulation in FINDPRO [78]. Merz and Strecker [79] solve a free boundary problem numerically by making assumptions for the boundary conditions. Another approach is to fit the logarithmic growth law to experimental data [80]. Ling et al. [81] extend the power-law model to a semiempirical oxidation model.

Boundary element-based algorithms can be used to focus specifically on sidewall masked oxidation [82] or to solve steady slow viscous flow [83]. The probabilistic or Monte Carlo approach can also be used [84]. The majority of numerical models are incorporated into process simulators. Some historical examples of these simulators include composite (one-dimensional) [85], FINDPRO (two-dimensional) [77,78], SMART-P (three-dimensional) [86], DIOS (two-dimensional) [79], PREDICT-1 [80,81], ICECREM [81], and, of course, SUPREM-3 [80,81,87], although Stanford University no longer supports SUPREM modeling. The majority of current models are available as commercial solutions. Silvaco International provides a commercial model called Athena (www.silvaco.com). Synopsys (formerly AVANTi Corporation and prior to AVANTi technology modeling associates (TMA)) no longer develops or maintains older models such as SUPREM-3, but does provide support for process simulation (previously integrated software environment (ISE) software) as part of its technology computer-aided design (TCAD) package (www.synopsys.com).

9.4.2 Limitations of Models

Limitations of these numerical models vary depending on the numerical analysis method used or the underlying theory. Models, which focus on stress effects, such as bird's beak or local oxidation of silicon (LOCOS) or STI isolation can have large errors in results due to assumptions used. Viscous or visco-elastic flow requires assumptions or approximations of physical parameters such as L_{eff} [85]. In order to simplify analysis this type of parameter is fixed, whereas in reality, L_{eff} varies with temperature.

Oxidized nitrides or oxynitride have been a part of semiconductor processing for many years. The study of these films has only recently become intensive because of the requirement for ultra-thin (<3 nm) gate dielectric films. The majority of the models begins with the Deal–Grove oxidation model and modifies it to include the effects of nitrogen at the Si–SiO_2 or Si–SiON interface.

Nitrided oxides or nitridation of oxides is considered to be a variation of silicon oxidation. One model assumes that nitrogen from source gases, such as N_2O is believed to neutralize growth sites at the Si–SiO_2 interface and results in retardation of the oxidation process [87,88]. The concentration of growth sites available at the interface directly affects the process. This model is derived from a modified Deal–Grove model. Another model investigates growth kinetics of oxidation of silicon in N_2O ambients by studying the effects of reaction energy for the formation of SiO_xN_y layers [75]. Oxidation of nitride layers is proposed to occur by diffusion and reaction of interstitial oxygen through the film to the Si–SiON interface [89]. The nitrogen is slowly displaced as the oxidation continues.

Low pressure chemical vapor deposition (LPCVD) of silicon nitride has been used in silicon processing since the early stages of semiconductor processing, but the level of effort for modeling does not match that of oxidation. There is no widely accepted model for LPCVD nitride using dichlorosilane (DCS) and NH_3, although models for plasma enhanced chemical vapor deposition (PECVD) or rapid thermal nitridation (RTN) or remote PECVD nitride have been proposed. Ogata et al. have completed a kinetic study of nitride films using DCS and NH_3 [90]. Film thickness profiles are estimated using gas phase and surface reaction rates of these gases. Another analysis is reported by Pigeon et al. using kinetics of particles to study the process [91]. A mathematical model for nitride films using DCS and NH_3 is proposed to study the effects of process and film uniformity on thickness [92]. This model is compared to experimental results and may provide optimization of the film growth conditions.

9.5 Metrology of Dielectric Films

Thin dielectric films pose a challenge to metrology on several fronts. Interfacial oxides between top layer(s) and bottom substrate are variable in quality, composition, and thickness causing errors from modeled behavior. Slight changes in composition of the thin film itself may have significant effects in the metrology of that film. Often, there is no universal standard for the film or known constants. Thickness measurement tools use models, which may not be applicable outside a defined range. Very thin films have variable stoichiometry across a wafer or inadequate models.

Semiconductor manufacturing fabrication facilities have partly solved these challenges using techniques other than physical thickness. After all, the only criteria the device is sensitive to is the electrical thickness. Techniques, such as capacitance–voltage or current–voltage (C–V) measurement can identify electrical thickness which has given certain information.

Unfortunately, physical thickness measurement is still needed for processing control in the fab. Convoluting factors, such as surface contamination, metal or contact film variations, etc. reduce the usefulness of information obtained from electrical measurements for process control.

The tables below provide some guidance as to techniques currently available for thin film characterization and measurement. Key references are provided for additional detail and information. In addition, the chapter on Metrology in this handbook is highly recommended.

9.5.1 Film Thickness Measurement Techniques

There are a variety of methods to measure the film thickness of silicon oxide. Each of these methods has limitations and may not be suitable for routine measurements. The following provides a qualitative description of some of these methods and associated limitations. Table 9.4 is a quick reference guide to current analytical, electrical, and qualitative techniques for film characterization. Table 9.5 provides a list of common dielectric films and their dielectric constants. Table 9.6 defines currently available national institute of standards (NIST).

9.5.1.1 Ellipsometry

Ellipsometry is a very powerful and accurate optical method to measure the thickness of silicon oxide films [93–95]. The technique uses plane polarized monochromatic light to illuminate the oxidized silicon surface at an angle. The light is reflected from both the silicon substrate and the oxide surface. The direct outputs of an ellipsometric measurement are the angles ψ and ϕ, which are related to the refractive index of the silicon substrate, the refractive index of the silicon oxide, and the thickness of the film. The refractive index for thick silicon oxide is usually assumed to be 1.462 for a wavelength of 632.8 nm [96]. This technique requires that the thickness is known to be within approximately 250 nm since the measured quantities are periodic functions of this thickness. Thicknesses down to the very thin (<5 nm) regime may be measured. However, inaccuracies in this regime may be present. The technique requires a known refractive index of the oxide film and assumes that the refractive index does not vary with thickness. More complicated methods, such as spectral or multiple wavelength ellipsometry [96,97] or multiple angle ellipsometry are now being used to improve the accuracy of these measurements to <1 nm [97,98].

9.5.1.2 Optical Interference

Direct optical interference methods may be used to measure the oxide thickness from several hundred to thousands of Angstroms. The technique is based on the interference of light reflected from the silicon substrate and oxide surface. The thickness is determined by finding the maxima and/or minima of the reflected light, either as a function of incident angle at constant wavelength [93,99] or as a function of wavelength at constant angle [100,101]. The variable wavelength technique is more common because the thickness of thinner films can be determined. Usually the maxima and/or minima are found at many wavelengths and compared to a fringe chart to determine the thickness [101]. The precision of the measurement, assuming a known constant refractive index, is approximately ± 2.5 nm in the 50–1000 nm range and is limited by the ability of the system to locate the maxima and minima [102].

9.5.1.3 High Resolution Transmission Electron Microscopy

Film thickness can be determined by using a micrograph obtained by high resolution transmission electron microscopy (HRTEM) [103–105]. The sample requires the tedious preparation of slicing the wafer, mounting the slice to a 3-mm transmission electron micrograph (TEM) grid, mechanical polishing and ion-milling it to approximately less than 50-nm thick [106,107]. A recent addition to the sample preparation is the use of focused ion beam (FIB) instruments [108], which uses ions to mill away both sides of the area of interest leaving a thin membrane for TEM. Under optimum sample

TABLE 9.4 Metrology Options Table for Analytical, Electrical, and Qualitative Techniques for Film Characterization

	Physical Properties						Chemical Properties					Electrical Properties						
Techniques	Physical Thickness—Single Layer	Physical Thickness—Film Stack	Pore Size/Distribution	Stress/Strain (Film/Interface)	Roughness (Interface/Film)	Density	Film Adhesion	Chemical Contaminant	Elemental Composition	Depth Dependent Elemental Composition	Chemical State	Electrical Thickness	Junction Depth	Dopant Concentration	Carrier Lifetime	Flatband Voltage	References	Issues/Limitations
Material Characterization																		
Raman				X	X				SiGe possibly								[2, p. 31]	
Scanning electron microscopy (SEM)	X	X															[2, p. 33]	Film modeling
Energy dispersive x-ray spectroscopy (EDX)	X	X															[2, p. 33]	
X-ray photoelectron spectroscopy (XPS)	X	X						X	X		X						[2, p. 25]	< ~50 A top film; large spot size
XPS at different incident angles	X	X							X	X	X						[2, p. 25]	
X-ray reflectivity/grazing incidence X-ray reflectivity (GI–XRR)	X	X	X		X												[2, p. 43]	Requires modeling of interface
X-ray diffraction (XRD)		Yes for some high resolution setups for repeating layer structures		X													[2, p. 31]	
X-ray fluorescence (XRF)	X	X															[2, p. 34]	Requires known sample density
Photoreflectance			X	X					SiGe								[2, p. 31]	
Interferometry				X													[2, p. 31]	
Secondary ion mass spectroscopy (SIMS)								ToF-SIMS		Depth profile			X	X	X		[2, p. 28]	
Electron beam induced XRF									X									

Ion beam sputtering										X						[2, p. 25]	
Electron micro-probe analysis	X	X						X			X			X		[2, p. 26]	Known physical thickness; large beam size
Laser induced surface acoustic wave analysis	X	X			Density dependent		Some info X									[2, p. 35]	Single layer films; requires known sample density
Laser acoustic technology In Line Process Control	X	X		X		X	X									[2, p. 39]	
Multiple angle ellipsometry	X	X														[2, p. 16]	Requires known optical constants to generate model
Single wavelength ellipsometry	X	X															
Spectroscopic reflectometers	X	X															
Spectroscopic ellipsometry	X	X						SiGe can determine Ge								[2, p. 16]	Optical constants for non-SiO_2 films vary locally; constants not well known [p. 20]
Minority carrier life time (MCLT)				X										X			
Gate oxide integrity (GOI) Electrical Characterization											X						
Non-contact corona based C–V								X			X	X	X	X	X	[2, p. 24]	
I–V											X			X		[2, p. 16]	Films $< \sim 20$ A, non-uniform doping [p. 22]
											X			X		[2, p. 24]	

Source: From Diebold, *Handbook of Semiconductor Manufacturing Technology*, 2005.

TABLE 9.5 Common Dielectric Films and Constants

Films	Dielectric Constant
SiO_2	3.9
Si_3N_4	7.8
ZrO_2	~25
HfO_2	~22
$ZrSiO_4$	~15
$HfSiO_4$	~13
HfON	20-25
HfSiON	3.9 to >25
Ta_2O_5	~25

Source: From Ino, T. et al., *Extended Abstracts of the 2005 International Conference on Solid State Devices and Materials*, 230-1, 2005.

thickness with limited surface roughness and microscope imaging conditions, errors on the order of 0.2 nm are generally achievable [96]. Unlike other techniques, no physical values such as refractive index or dielectric constant are needed.

9.5.1.4 Electrical Methods

Electrical measurements may also be used to determine the thickness of the oxide in a metal oxide semiconductor (MOS) capacitor. The MOS capacitor may be fabricated as part of a standard test chip with standard technology (e.g., polysilicon gate, STI or LOCOS, etc.). A MOS capacitor may also be formed by depositing metal dots on the oxide or by using a drop of mercury in a mercury probe configuration. However, the deposited metal deposits may spike into the oxide especially at high temperatures. The most common measurement method is the C–V technique [102]. In this technique, either a low- or high-frequency C–V curve is measured. In the classical approximation, the capacitance in strong accumulation (C_{acc}) or strong inversion at low frequency (C_{inv}) is equal to the oxide capacitance (C_{ox}), which is related to the oxide thickness as

$$C_{acc} = C_{inv} = C_{ox} = \frac{\varepsilon_{ox}}{X_o} A_g \tag{9.32}$$

where ε_{ox} is the static dielectric constant of the insulator and A_g is the area of the gate (capacitor). The dielectric constant for thick silicon oxide is usually assumed to be $3.9\varepsilon_0$. There are many assumptions with this model. A single homogeneous film with constant static dielectric constant is assumed [96]. This model does not account for tunneling or leakage currents that may be present in thin oxides [102,109]. This model also does not account for the finite semiconductor capacitance associated with the semiconductor surface charge. For thin oxides, the semiconductor surface charge has a finite thickness of the order of 3–10 nm, which must be taken into account to properly find the oxide capacitance [96,110]. Also, a voltage drop may be present across the gate (polysilicon depletion), which must be taken into account [96,110]. Other techniques based on capacitance have also been suggested to alleviate some of these problems [96,111,112].

Another electrical method to determine the oxide thickness is the use of Fowler–Nordheim tunneling. The I–V characteristics of a MOS capacitor can be measured and the results compared to the simple Fowler–Nordheim theory based on tunneling across a triangular energy barrier [96]. This method does not consider deviations from the ideal case, such as the effects of band-bending, quantization of electrons, thermal broadening of energy levels or surface roughness. Furthermore, the effective mass of the electron in the oxide and the oxide bandgap (electron barrier height) must be known and must not deviate across the film. Fowler–Nordheim tunneling may also be used to determine the thickness by analyzing the oscillations in the current due to the constructive interference of the incoming and reflected electron wave [113]. These limitations are addressed in J. Hauser's C–V model [114].

TABLE 9.6 Available NIST Standards

Depth Profiling		Film Thickness		Resistance			
Implant	Concentration	Film	Thickness (nm)	Spreading Resistance (ohm cm)	Nominal Bulk Resistivity (ohms-cm)	Nominal Sheet Resistance (ohms)	Sheet Resistance
			NIST Standards (Government)				
P implant in Si	9.58×10^{14}	SiO_2–Si	50				
As in Si	7.330×10^{14}		100	0.001–200			
NiCr thin film	Cr: 41.3 μg/cm^2; Ni: 49.4 μg/cm^2		25	0.01			
B in Si	1.018×10^{15}		14	0.1			
				1			
				10			
				25			
				100			
				200			
			VLSI Standards (Commercial)				
		SiO_2/Si	1010		0.01	0.1	
			675		0.03	0.4	
			400		0.1	1.4	
			200		0.03	4	
			50		0.88	12	
			25		3	41	
			12		10	138	
			7.5		30	414	
			4.5		60	828	
			2		88	1214	
		Low pressure chemical vapor deposition (LPCVD) Si_3N_4 on Si	200				3–6 ohms on 381 μm bulk
			120				20–25 ohms on 381 μm bulk
			90				
			20				

9.5.1.5 Other Methods

A variety of other less-common methods can also be used to determine thickness. A mechanical surface profiler or atomic force microscope (AFM) can be used to measure thickness from approximately 10 to 5000 nm by mechanically scanning a stylus over a step etched between the silicon and oxide surfaces. The weight gain method uses the increase in weight due to oxidation measured by a vacuum microbalance and an assumed density of silicon oxide to determine the film thickness. A variety of analytical techniques may also provide a measurement of thickness, such as x-ray fluorescence (XRF), x-ray photoelectron spectroscopy (XPS), Rutherford backscattering spectrometry (RBS), and medium energy ion spectrometry (MEIS). However, these techniques usually require a complicated parameter extraction technique and some are time-consuming. As new multi-component materials are introduced into the CMOS flow, the above techniques are becoming essential in controlling the material properties both inline and offline. Please read the Metrology chapter in this handbook for a more in-depth understanding of the individual techniques.

9.5.2 Gate Oxide Integrity (GOI) and Measurement Techniques

To ensure the reliable operation of MOS devices over its lifetime, the integrity of the silicon oxide gate dielectric is crucial. Issues with gate dielectric reliability fall into the two main categories of dielectric breakdown and hot-carrier degradation.

9.5.2.1 Oxide Breakdown

The catastrophic breakdown of a dielectric is generally believed to be a two-stage process [115–117]. First, the dielectric is degraded over time due to the passing of current or high electric fields. This degradation takes the form of defects in the oxide, which form a localized conductive path through the oxide as a function of time. The second stage is the breakdown process itself in which the oxide is shorted and a very large current flows. There are numerous physical models attempting to describe oxide degradation including lattice damage models [91], trap creation models [118], hole-induced degradation models [117,119,120], electron trapping models [121], interface state generation models [122], and the resonant tunneling model [123]. Although the physical basis for oxide degradation is still under debate, definitions and practices have been developed to describe and test the integrity of the gate oxide.

There are several methods used to characterize the electrical breakdown of oxides [115]. The ramped voltage breakdown test uses a linearly ramped voltage applied to a MOS capacitor until the oxide breaks down and the current drastically increases. This voltage at which breakdown occurs (V_{bd}) is sometimes referred to as the oxide dielectric strength. The other main test is that of charge- or time-to-breakdown at either a constant voltage or constant current. In these tests, a large constant voltage (current) is applied to a MOS capacitor until breakdown occurs in the form of a current increase (voltage decrease), marking the time to breakdown (t_{bd}). The charge-to-breakdown (Q_{bd}) is determined from

$$Q_{bd} = \int_0^{bd} J_g(t)\mathrm{d}t \tag{9.33}$$

where $J_g(t)$ is the capacitor gate current as a function of time. Equation 9.33 reduces to the following for constant current stressing

$$Q_{bd} = J_g \times t_{bd}. \tag{9.34}$$

The above measured parameters are usually taken on a large number of samples. It has been found that all oxides from a given process do not have the same breakdown characteristic, but instead have a distribution of values. As reviewed in Ref. [115], analyses of data using various statistical functions have indicated that the types of breakdown or oxide failures fall into three groups. The first group of oxide failures occurs instantly upon application of a small bias. These failures are generally due to gross defects

in the oxide, which cause a short of the dielectric. The second group of oxide failures occurs under intermediate stresses; do not instantly short, but cause early failures in integrated circuits. These failures are believed to be due to weak spots or defects in the oxide. The final group of oxide failures is considered to be due to intrinsic properties of silicon oxide. These failures are believed to occur in defect-free oxides and these oxides can withstand the highest stressing conditions. These final two groups are many times quantitatively modeled to determine, if the device would have a long enough lifetime (usually considered to be 10 years) under normal operating conditions [124].

9.5.2.2 Hot-Carrier Degradation

As the metal oxide semiconductor field effect transistor (MOSFET) device dimensions are scaled, the electric fields found in the device become increasingly high resulting in reliability problems. Specifically, the larger electric fields found in the channel result in impact ionization and the creation of hot-carriers [125]. In an *n*-channel MOSFET, most of the generated electrons enter the drain and most of the holes enter the substrate resulting in a measured substrate current. Some of these hot-carriers, however, can be injected into the gate oxide resulting in the degradation of the oxide. This degradation can increase oxide charge and interface state density, which can then cause Vt instability and current drive degradation [124,126]. Since holes are cooler than electrons, hot-carrier degradation in *p*-channel MOSFETs is much less than that in *n*-channel MOSFETs [127]. The location of damage is found to be mainly above the drain as the electric fields are highest in this region.

Several device parameters are usually monitored to characterize the impact of hot-carrier degradation. These include Vt shift, drive current reduction, transconductance degradation, and increase in interface state density (proportional to shift in subthreshold swing). The lifetime of a device under given biasing conditions is then chosen based on a given change in one or more of the above parameters. The ranges of these criteria vary depending on application, but usually are in the range of 10–100 mV Vt shifts and percent changes in the other parameters between 3 and 10% [115,126]. To assess the impact of hot-carrier degradation at real MOSFET operating voltages, the stress level is increased to accelerate the failure time. A model is then used to extrapolate the results to normal operating conditions [115,126]. A device lifetime of at least 10 years is usually chosen as the design criteria for reliable operation under these normal operating conditions.

9.5.3 Oxide Charge, Bulk Trap and Interface Trap Measurements

The silicon–silicon dioxide system contains various charges and traps. The charges and traps can modify device operation in many ways, such as changing the Vt, degrading current drive, or degrading device reliability, and yield. The traps and charges are generally classified into the four types of fixed charge, mobile ionic charge, oxide traps, and interface traps [128]. The following provides a brief review of each of the charges and of associated measurement techniques [102,128,129].

The fixed charge density (Q_f) is located within approximately 3 nm or less for very thin dielectrics of the interface is usually considered positive and is related to the oxidation and annealing conditions as well as surface orientation. The most widely used method to determine fixed charge is the C–V technique. The flatband voltage is first determined from the C–V method. Assuming that Q_f is a sheet of charge located at the Si–SiO_2 interface and that it dominates the total oxide charge Q_f can be found as,

$$Q_f = C_{ox}(-V_{fb} + \phi_{ms}). \tag{9.35}$$

Q_f may also be found from the slope of a Q_f vs. x_0 curve, since C_{ox} is related to x_0 through Equation 9.32. This technique is tedious because it requires measurements on capacitors with different oxide thickness. However, there is less error associated with this technique because it is independent of ϕ_{ms}. A tapered oxide technique formed by etching the oxide can be used to shorten the time to obtain the different oxide thicknesses needed for these measurements.

The mobile ionic charge (Q_m) is attributed to ionic impurities, such as Na^+, Li^+, and K^+ as well as to negative ions and heavy metals. These ions are mobile, even at room temperature, when an electric field is

present. The mobile ionic charge is determined by heating the sample, drifting the charge to one interface by applying an electric field, cooling the sample, and measuring C–V. The process is repeated again with an electric field of opposite polarity. The mobile charge is then determined as

$$Q_m = -\Delta V_{fb} C_{ox}. \tag{9.36}$$

If the C–V curves become distorted due to interface traps induced by the mobile charge drift, ΔV_{fb} may be difficult to determine resulting in errors in Q_m. The mobile charge density may also be determined through the triangular voltage sweep (TVS) method, which involves analyzing the capacitor current induced by a voltage ramp at both room and elevated temperature. Triangular voltage sweep is not affected by induced interface traps.

The bulk of the oxide can contain both charged and neutral defects that are capable of trapping electrons or holes. These bulk traps may trap free charge induced in the oxide by ionizing radiation or high current flow, thus affecting device performance and lifetime. The procedure for measuring the electrically active bulk oxide traps is to measure an electrical parameter such as flatband voltage, introduce a specific number of electrons or holes into the oxide, re-measure the electrical parameter and relate the change in the electrical parameter to the trapped charge density. There are many methods to introduce electrons or holes into the oxide. Some commonly used techniques are internal photoemission with a high energy (3–6 eV) light, avalanche injection of carriers using a triangular voltage pulse, optically assisted injection using a MOSFET, and normal microscope light, Fowler–Nordheim tunneling and injection using a low energy electron beam. Several techniques may be used to measure the electrical parameter. The photo C–V technique may be used to measure the number of trapped charges, the charge centroid and perhaps the charge distribution. The flatband voltage from C–V or the Vt from a MOSFET may be used to determine the number of filled traps; however, the location of the charge must be determined separately.

Interface traps (fast or surface states) are located at the silicon–silicon dioxide interface and, unlike the other charges, are in direct electrical communication with the silicon. These states contribute energy levels over the silicon bandgap and are occupied based on the silicon surface potential. These defects may be reduced through the use of a low temperature ($\sim$450°C) forming gas (H_2, N_2) anneal. There are many techniques to measure interface trap density including quasi-static C–V, charge-pumping, and Deep Level Transient Spectroscopy.

9.5.4 Issues with Ultra-Thin Film Measurements

As gate dielectrics for CMOS technology scale to the ultra-thin regime ($<$2 nm), the issues associated with measuring these films becomes increasingly critical. A transition region exists between the crystalline silicon and bulk silicon oxide that extends anywhere from 0.55 to 3 nm from the silicon [95,129–132]. The transition region has been attributed to non-stoichiometric silicon oxide, strain, changes in bond angles, or differences in bonding arrangement. The structural differences in the transition region can change the properties of the film including refractive index, static dielectric constant, band gap, and effective mass. These property changes can affect many of the measurements described above. Surface roughness also plays a larger role in the measurements of these films, since many of the measurements assume an atomically smooth transition between silicon and silicon oxide.

Leakage and tunnel currents through the oxide will affect many metrology aspects including capacitance and conductance measurements, transistor properties, and theoretical assumptions associated with electrical measurements [109,133]. Effects associated with the MOS structure include inversion layer quantization and polysilicon gate depletion [96,110,134]. In inversion layer quantization, the high electric fields associated with the thin oxide MOS structure causes energy level splitting in the silicon. This quantum mechanical effect results in a shifting of the inversion layer electron centroid to approximately 0.4–0.5 nm below the surface, which effectively reduces the gate capacitance and inversion charge density. The polysilicon depletion effect, unless a metal gate electrode is used, also reduces the gate capacitance and inversion charge density for a given gate drive. Each of these effects must be considered

when electrically measuring a MOS device with ultra-thin oxide. Finally, issues associated with reliability may change for ultra-thin oxides and further study will be required to determine proper reliability metrology.

9.6 Gate Dielectrics

9.6.1 Plasma Nitrided Oxide

9.6.1.1 Role on Nitrogen

As devices are aggressively scaled, the gate dielectric thickness must be correspondingly decreased. The thinning of the gate dielectric has given rise to very high gate leakage current with significant implications on device reliability that cannot be ignored. Incorporation of nitrogen atoms in ultra-thin gate dielectrics with an equivalent oxide thickness (EOT) less than 3 nm has been shown to reduce gate leakage and boron diffusion from boron-doped p^+ poly-Si gate electrode to the channel region and to enhance reliability of MOS devices without sacrificing performance [135].

Monolayer-level nitrogen incorporation (Si_3–N bonding configuration; ~ 6–$8 \times 10^{14}\ cm^{-2}$) at the Si–$SiO_2$ interface reduced gate leakage by approximately tenfold compared to pure SiO_2 as shown in Figure 9.6 [136]. The incorporation of the nitrogen at Si–SiO_2 interface leads to reduced interfacial sub-oxide bonding [136], which defines a transition region between the Si substrate and bulk SiO_2, and that the modified interface structure results in reduction in gate tunneling currents by increasing tunneling barrier height [131]. Figure 9.7 shows schematic band diagrams for gate stacks with (a) sub-oxide and (b) nitrided interface. Modification of the conduction band offset for the nitrided interface results in the gate leakage current reduction.

Yang et al. [137] calculated gate leakage currents for various composition of bulk homogeneous SiON, which is represented by pseudo-binary expression, $(SiO_2)_x(Si_3N_4)_{1-x}$. The simulation assumes a linear

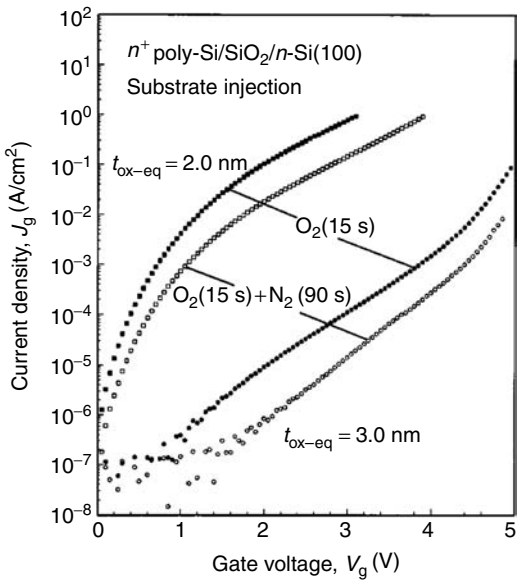


FIGURE 9.6 Gate leakage current density vs. gate voltage traces for 3.0- and 2.0-nm thick gate dielectrics on ⟨100⟩ silicon, demonstrating the effect of the monolayer of interfacial nitrogen (symbols in white) reducing tunneling currents compared to the pure oxide (symbols in black). (From Niimi, H. and Lucovsky, G., *J. Vac. Sci. Technol. B*, 17, 2610, 1999.)

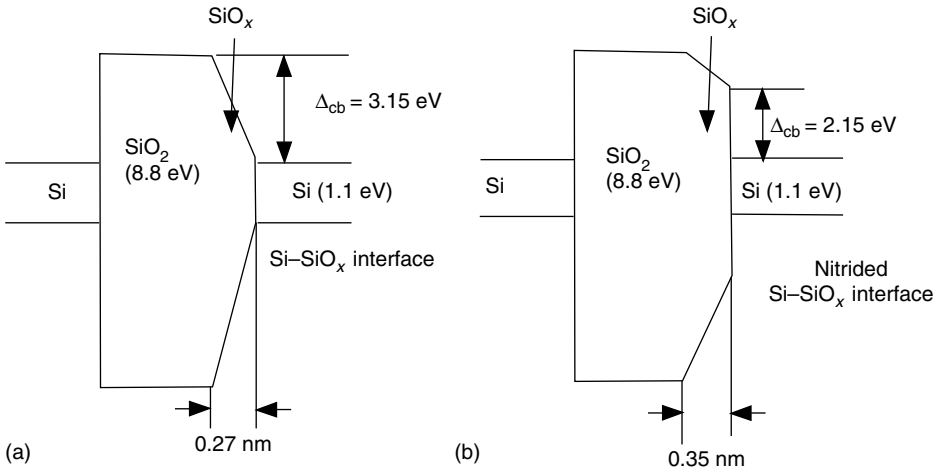


FIGURE 9.7 Schematic representation of band diagram for gate stack with (a) sub-oxide (SiO_x, $x < 2.0$) interfacial transition region and (b) nitrided interface. (From Yang, H. et al., *Electron Device Lett.*, 21, 76, 2000.)

variation of the dielectric constant, effective mass, and conduction band offset energy [137]. Figure 9.8 shows leakage current as a function of fraction of silicon nitride in SiON. The leakage current as a function of composition shows a minimum at a composition of about 60% of Si_3N_4. Ellis et al. [138] simulated the effects of a nitrogen distribution in an oxide for boron diffusion using the random walk method. In this model, the boron substitution at a Si site was blocked via the presence of Si–N bonds. The removal of boron diffusion pathways via the Si–N bonds results in the reduction of boron diffusivity. Ellis et al. [138] investigated three types of nitrogen profile in the oxide vs. the boron diffusivity: (1) nitrogen distributed uniformly throughout the oxide (as in uniform bulk oxynitride alloy), (2) nitrogen existed only near the Si–SiO_2 interface, and (3) nitrogen incorporated at the top surface of the oxide.

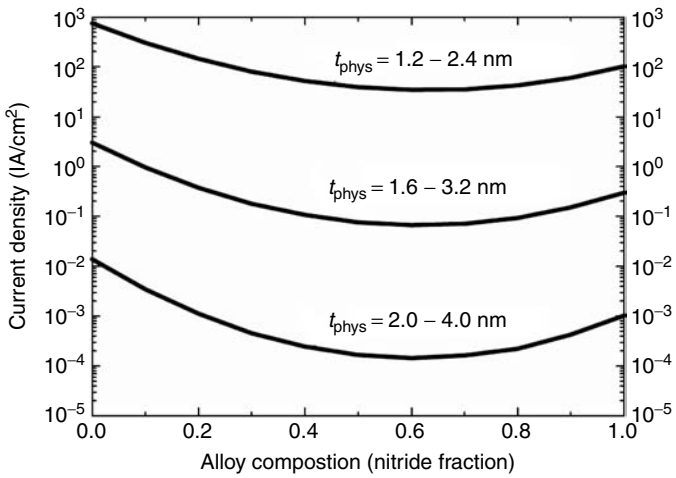


FIGURE 9.8 Calculation of tunneling current at $V = 1$ V for constant equivalent oxide thickness. It assumes linear variation of dielectrics constant, tunneling mass, and conduction band offset. (From Yang, H. and Lucovsky, G., *Technical Digest of IEEE International Electron Device Meeting*, 1999, 245.)

According to the simulation, if the integrated nitrogen concentration were fixed, top surface nitridation would be the most effective way to block the boron diffusion paths.

Therefore, controlled incorporation of nitrogen, especially content and location/profile in the ultra-thin gate dielectrics is the key to engineering gate dielectrics for future generation devices.

9.6.1.2 Plasma Nitrided Oxide

One of challenges for the thermal nitridation is an independent control of dielectric thickness, nitrogen content and/or location. Since oxidation and nitridation occur spontaneously, it is difficult to form sub-1.0 nm thick ultra-thin gate dielectrics with higher nitrogen concentration (more than 10 atm%). The maximum concentration of nitrogen incorporation is limited by thermodynamics.

Since Lucovsky et al. proposed plasma nitridation of the gate oxide [139], plasma nitrided oxide (PNO) process has been established as a standard method of forming ultra-thin gate dielectrics since 180 nm node [140–143].

The advantage of the plasma nitridation of SiO_2 is in its ability to control (1) the dielectric layer thickness and (2) the content and location and/or profile of nitrogen incorporation independently and precisely. In addition, since it is plasma, i.e., a non-thermal process, the nitrogen can be incorporated much more than the thermal processes and the profile, i.e., location can be carefully engineered. Plasma nitrided oxide can be created by (1) top surface nitridation, (2) homogeneous SiON formation, where the nitrogen is distributed uniformly through the oxide, and (3) incorporation of nitrogen at the Si–SiO_2 interface.

When a 3-nm SiO_2 film is exposed to a remote plasma of He–N_2 at 0.1 and 0.3 Torr, top surface nitridation is achieved, whereas higher pressure nitridation forms homogeneous SiON, as shown in Figure 9.9 [144]. Figure 9.10 illustrates the optical emission from a H_2–N_2 plasma as a function of process pressure [144]. The data indicates that ionic metastable nitrogen N_2^+ is responsible to the top surface nitridation and neutral nitrogen nitridation tends to form uniform SiON. The aforementioned study indicated that the nitridation mechanism changed from non-diffusive (charged particle assisted) forming top surface nitride layer to diffusive (presumably involving neutral nitrogen metastable species) forming uniform SiON as the process pressure was increased from 0.1 to 0.3 Torr. This is an example of how nitrogen profile engineering using plasma chemistry can be achieved.

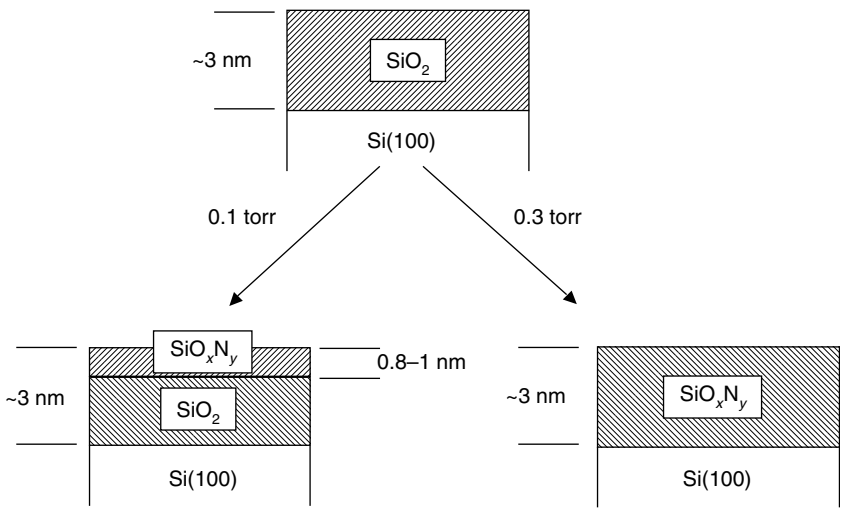


FIGURE 9.9 Schematic representation of the nitridation process and the resultant nitrogen distribution for He/N_2 plasma process at (a) 0.1 Torr and (b) 0.3 Torr. (From Niimi, H. et al., *J. Appl. Phys.*, 91, 48, 2002.)

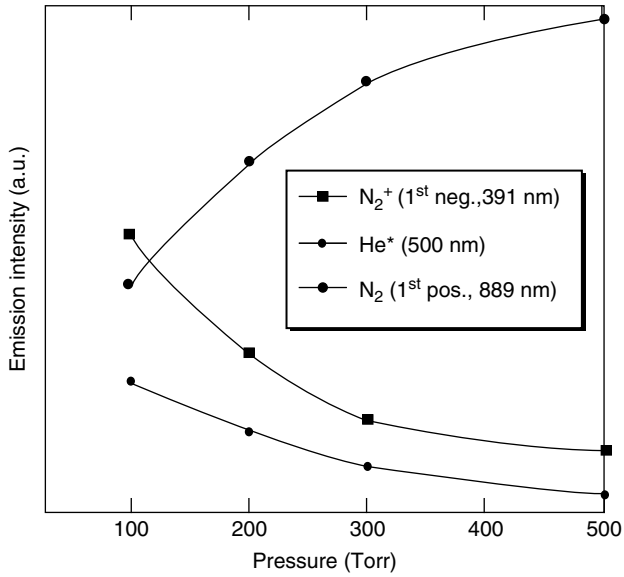


FIGURE 9.10 Optical emission intensities of N_2^+ (ionic nitrogen metastable, 391 nm), N_2 (neutral nitrogen metastable, 889 nm), and He (500 nm) as a function of pressure for He/N_2 plasma. (From Niimi, H. et al., *J. Appl. Phys.*, 91, 48, 2002.)

Plasma nitrided oxide is basically three step process: (1) oxidation of Si surface, (2) plasma nitridation of SiO_2, and (3) reoxidation of SiON. Figure 9.11 illustrates the PNO process sequence. The initial step, oxidation defines the physical thickness gate dielectrics. The second step, plasma nitridation, incorporates nitrogen atoms into the SiO_2 and the final step, reoxidation, removes plasma damage. Additionally the reoxidation step densifies the gate dielectric, minimizes unfavorable dangling bonds, decreases the nitrogen concentration at Si–SiON interface and as a result improves device performance through mobility enhancement and reliability.

Oxidation can be performed either in a furnace (batch) or single wafer chamber. For high quality ultrathin oxide, for example, use Applied Materials (AMAT), ISSG or TEL, and water vapor generator (WVG) oxidation reactors. Both processes are capable of forming sub-2 nm silicon dioxide films.

Plasma nitridation can be performed using either radio frequency (RF) or microwave-plasma sources. The most widely used tools in the industry are AMAT's RF-type decoupled plasma nitridation (DPN) plasma and TEL microwave-type slot plate antenna (SPA) plasma technologies. Nitrogen gas is mainly used as the nitrogen source, but nitric oxide (NO), nitrous oxide (N_2O), and/or ammonia (NH_3) can also

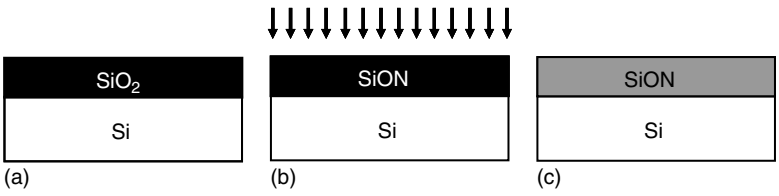


FIGURE 9.11 Plasma nitrided oxide (PNO) process 3-step sequence: (a) oxidation, defines physical gate dielectrics thickness; (b) Plasma nitridation, incorporates nitrogen atoms into the SiO_2; and (c) Reoxidation (ReO_x) removes plasma damage via the nitridation step, densities the gate dielectrics, and minimizes unfavorable dangling bonds.

be used for plasma nitridation process. In addition to the nitrogen source gas, inert gases such as helium (He), argon (Ar), krypton (Kr), and xenon (Xe) have also been used. For example, the use of helium in nitrogen plasma provides an additional kinetic path, i.e., penning ionization and thus increases rate of nitridation [144].

9.6.1.3 Reliability

An originally unexpected benefit of the PNO is the extension of the reliability scaling limit of SiO_2-based gate dielectrics [145]. SiON with a uniform nitrogen profile overcame the reliability scaling limitations of SiO_2 via gate leakage reduction, Weibull slope improvement, and voltage acceleration factor enhancement [143]. However, there are also some disadvantages. Electron spin resonance studies have shown that the nitridation of oxide films create bridging nitrogen center precursors [146]. These bridging nitrogen center precursors at the Si–SiO_2 interface, in general, have been found to degrade negative bias temperature instability (NBTI). The bridging nitrogen center reduction can be achieved, for example, by plasma nitridation engineering to have peak nitrogen concentration away from the interface [146] and post-plasma nitridation reoxidation optimization [147,148].

9.6.1.4 Sub-1.0 nm SiON

In order to scale SiON to below 1.0 nm EOT with low leakage, it is necessary to increase the nitrogen concentration. The precise nitrogen concentration needs to be determined experimentally, but it is close to that calculated by Yang and Lucovsky [137]. In addition, since bonding constraint theory suggests that Si_3N_4 cannot be directly substituted for SiO_2 at the Si substrate [149,150], the composition at the Si–SiON interface must be carefully controlled. According to the bonding constraint theory, which established criteria for formation of low defect density glasses originally, the average number of bonds ($N_{av} \sim 3$) represents a criterion between low defect density and increasingly defect materials [151]. The average coordination at the Si–Si_3N_4 interface has an over-constrained bonding configuration ($N_{av} \sim 3.5$) with a significantly higher defect concentration. If an ultra-thin SiO_2 layer (~ 0.5 nm) is formed between the Si substrate and Si_3N_4, the average bonding coordination is reduced down to 3.0, and is essentially the same as for the silicon–silicon dioxide interfaces leading to significantly improved electrical performance. Therefore, interface engineering is critical to achieve higher device performance. It is especially true for higher nitrogen concentration SiON.

Recently, plasma nitridation of Si surface has been demonstrated to achieve sub-1.0 nm gate dielectrics with drastically leakage reduction for 45 nm and beyond [152,153]. Both groups studied optimized post-plasma nitridation reoxidation processing to improve device performance. The SiO_2 interface layer enhances device performance and relatively high nitrogen concentration "bulk" silicon oxynitride layer reduces the gate leakage. If one could incorporate a mono-layer level of nitrogen at silicon–silicon dioxide interface the leakage could be reduced even further.

9.6.2 High-*k* Gate Dielectrics

9.6.2.1 Introduction

As discussed in the previous section, gate leakage is a major issue in SiO_2 and SiON as the EOT is scaled near 1 nm for high performance devices and below about 1.5 nm for low power devices. Silicon oxynitride has been extended beyond the point, where the industry originally thought could be achieved and is still being optimized to meet low leakage requirements; however, it is approaching a point where even the optimized [152] dielectrics have high leakage and potentially low mobility due to nitrogen incorporation at the interface for EOTs lower than 1 nm.

High-*k* gate dielectric is an obvious approach to solve the gate leakage problem. High-*k* materials have been investigated for nearly a decade, but none have been introduced into products according to the latest publications. Researchers in the semiconductor industry, universities and research organizations have investigated many high-*k* materials in an attempt to replace SiON. Among them are Al_2O_3, Y_2O_3, La_2O_3, HfO_2, ZrO_2, and their aluminates and silicates including nitrided silicates. These materials are

thermodynamically stable in contact with Si and have adequate band offsets. Many other materials with much higher dielectric constants like the perovskites, TiO_2, and Ta_2O_5 have also been investigated, but these have much lower band offsets and are unlikely to be used as gate dielectrics [154]. Among the materials with high-band offsets, it was found that the nitrided silicates specifically hafnium silicon oxynitride (HfSiON) [155–157] are physically and electrically stable enough to be considered. Hafnium oxide (HfO_2) has a dielectric constant ranging from about 18 to 22 [158], and HfSiON has been reported to have a dielectric constant of up to 24 [159]. The dielectric constant of HfSiON can be varied by changing the hafnium to silicon ratio and/or the oxygen to nitrogen ratio. Hafnium-based gate dielectric compounds also have high band offsets, $\Delta E_{CB} \geq 1.5$ eV and $\Delta E_{VB} \geq 3$ eV that make them especially useful for MOS devices [154]. The aluminates also have high band offsets and might be better suited, if only the band offset and band gap were the critical parameters. However, it has been found that aluminum-based dielectrics have a lower crystallization temperature than the nitrided silicates, and reduced mobility as a result of aluminum diffusion into the channel [160].

The flexibility of changing the dielectric constant by changing the composition of the gate dielectric, as is the case for at least ternary compounds, can also introduce variation in the equivalent oxide thickness that will have to be addressed by optimizing the deposition and post-treatment processes. The simplest solution, of course would be to use simple binary oxides, such as HfO_2 since its composition, in this case, stoichiometry is more easily controlled. However, HfO_2 has many undesirable properties, such as low crystallization temperature, gives rise to low channel mobility [161], has a high charge trap density [162], is oxygen permeable as are most "ionic" metal oxide compounds, etc. Hafnium silicon oxynitride [155], on the other hand has been found to be amorphous over a wide composition range and devices have shown the highest thermal and electrical stability among the high-*k* gate dielectrics investigated to date. Furthermore, MOS devices using HfSiON gate dielectrics have shown the highest channel mobility [163,164], and can be scaled to EOT much lower than 1 nm even for polysilicon gated devices [165,166].

One of the primary reasons that high-*k* gate dielectrics have not been adopted into production yet is the flat band offset and gate depletion observed principally in pMOS devices, when doped poly silicon is used as a gate electrode. The flat band voltage offset has been attributed to Fermi level pinning, but not all of the published data is in total agreement. It is likely that the offset is a combination of high-*k*/poly silicon reactions leading to the so called Fermi level pinning [167], native defects such as vacancies [168], and/or dopant effects [169] are responsible. As a result of the difficulty in setting the Vt with poly silicon electrodes, metal gate electrodes are being investigated. The hope is that metal electrodes can solve both the high gate depletion as well as the flat band voltage offset. There is sufficient published data now, which indicates that gate depletion is significantly reduced when metal gates are used with high-*k* dielectrics. However, while there are reports that the Vt can be controlled using metal gates in a CMOS flow [170], there are no reports from integrated circuit (IC) manufacturers that suggest that high-*k* with metal gates have been introduced in production flows and have the required low controllable Vts.

9.6.2.2 Deposition

Hafnium-based gate dielectrics have been deposited mostly by physical vapor deposition (PVD), chemical vapor deposition (CVD), and atomic layer deposition (ALD). Physical vapor deposition has been successfully used to perform the initial pioneering deposition of hafnium silicon oxide (HfSiO) [171] and hafnium silicon oxynitride (HfSiON) [155,166,172]. The initial work performed using PVD HfSiON created the critical data that formed the basis for the direction; the industry has taken on high-*k* gate dielectrics over the last few years. Today the hafnium-based compounds are now being deposited by ALD and CVD, because of among other reasons compositional control and future potential conformality requirements. Atomic layer deposition has been successfully used to deposit HfO_2 using hafnium tetrachloride ($HfCl_4$) [173], and more recently ALD has also been used to deposit HfSiO using metal organic precursors such as tetrakis ethylmethylamido silicon (TEMASi) and tetrakis ethylmethylamido hafnium (TEMAHf) [174]. Chemical vapor deposition processes, on the other hand have used predominantly metal organic precursors such as a combination of tetrakis ethylamido hafnium (TDEAHf) and tetrakis methylamido silicon (TDMAS) [175], and of hafnium (IV) *tert*-butoxide

(HTB) and tetraethoxysilane (TEOS) [176] or disilane (Si_2H_6). These precursors have been selected after significant investigations that were targeted at minimizing SiO_2 interfacial growth during deposition, composition control, particle performance, and device property optimization to name a few.

As described in the PNO section, nitridation is necessary in order to reduce the leakage of SiO_2 and to block boron. In the case of HfSiO, Visokay et al. [177] found that nitridation eliminated phase separation and crystallization in the temperature regime used for conventional CMOS flows in addition to boron blocking. This key finding has been critical to the progress of high-*k* gate dielectrics. According to the latest publications, all HfSiONs with more than a few percent nitrogen are produced by post-deposition processes using either ammonia or plasma nitridation. The selection of the nitridation process may depend on the required nitrogen concentration and/or desired nitrogen profile.

Most of the equipment suppliers with high-*k* gate dielectric programs have demonstrated the capability to deposit both HfO_2 and HfSiO with subsequent nitridation either thermal annealing in NH_3 or plasma nitridation.

9.6.2.3 Physical Properties of HfO_2 and HfSiON

One of the key properties of SiO_2 and SiON is their amorphous nature even after annealing at temperatures well above the conventional CMOS processing temperatures. Hafnium oxide (HfO_2) and HfSiO, on the other hand have much lower crystallization temperatures. Figure 9.12 shows a cross-sectional TEM of HfO_2 that clearly shows ordering in the thin film after polysilicon deposition at about 700°C. Hafnium silicon oxide, on the other hand, in addition to crystallizing at temperatures lower than the maximum CMOS processing temperature shows phase separation as it is annealed at these temperatures as shown in Figure 9.13. Crystalline gate dielectrics, especially polycrystalline films will most likely have point defects and grain boundaries that are expected to place significant limitations on the reliability of the gate dielectric as reported by Yamaguchi et al. [178]. As already mentioned above, the addition of nitrogen to HfSiO to form HfSiON stabilizes the structure such that the material remains amorphous up to about 1100°C, for the composition reported (SiO_2 rich). The crystallization temperature is dependent of course, on the material composition; because of the structural properties and the initial encouraging electrical and device properties, the discovery of sought after properties of HfSiON provided the much needed focus on a narrower set of materials.

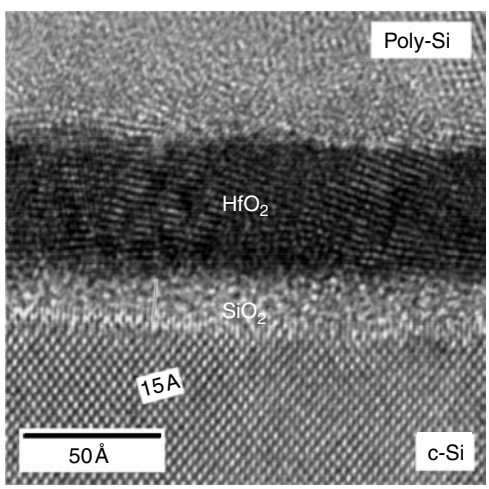


FIGURE 9.12 Transmission electron micrograph (TEM) of polySi/HfO_2/SiO_2/Si showing where the poly Si was deposited at approximately 700°C.

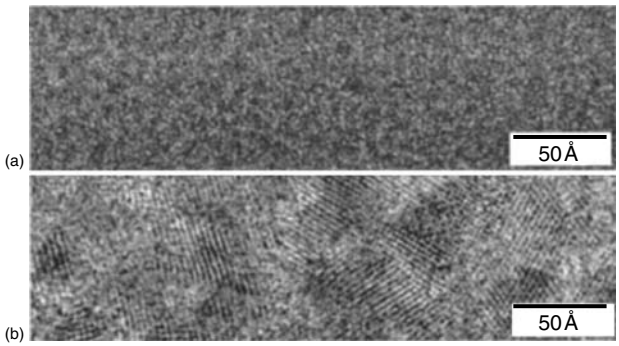


FIGURE 9.13 Cross-section high resolution transmission electron microscopy (HRTEM) images of thick approximately 100 Å HfSiON and HfSiO films after annealing for 60 s in N_2. (a) HfSiON annealed at 1100°C and (b) HfSiO annealed at 1000°C.

Figure 9.14 shows a TEM, a cross-section of polySi/HfSiON/SiO_2/Si after annealing at 1100°C for 15 s. The material is not crystalline after the high temperature annealing and does not seem to have reacted with the poly silicon electrode. The physical and electrical properties of HfSiON reported in the literature to date are consistent regardless of deposition technique and nitridation technique used. It is important, however, as in the case of SiON to control the composition gradient. The reproducibility of the HfSiON properties by different groups using different processes suggests that the observed properties are fundamental and are not a result of extrinsic effects related to the deposition processes or other factors. The amorphous nature of HfSiON has also been supported by x-ray diffraction and extended x-ray fine absorption structure (EXFAS) analysis [179].

Figure 9.15 shows the dielectric constant of HfSiON as a function of composition along with a calculation of the dielectric constant of a dielectric stack formed of thin SiO_2 and high-k of varying dielectric constant. The dielectric constant of the quaternary compound is higher than that of HfO_2, which is expected to allow EOT scaling of HfSiON to the physical scaling limits of gate dielectrics on Si channels based on simple structural and thermodynamic arguments. Scaling of EOTs below 0.6 nm will be very challenging regardless of high-k material employed because of the interfacial silicon oxide thickness.

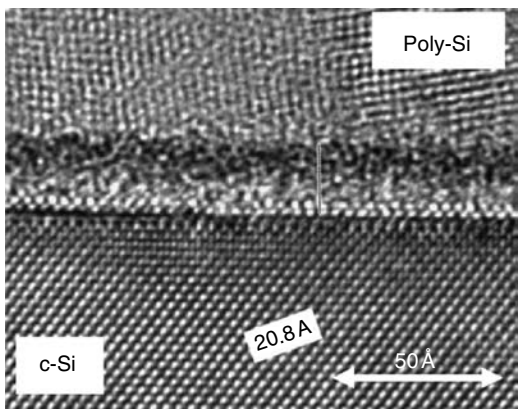


FIGURE 9.14 TEM of polySi/HfSiON/Si after annealing at 1100°C in a nitrogen ambient for 15 s.

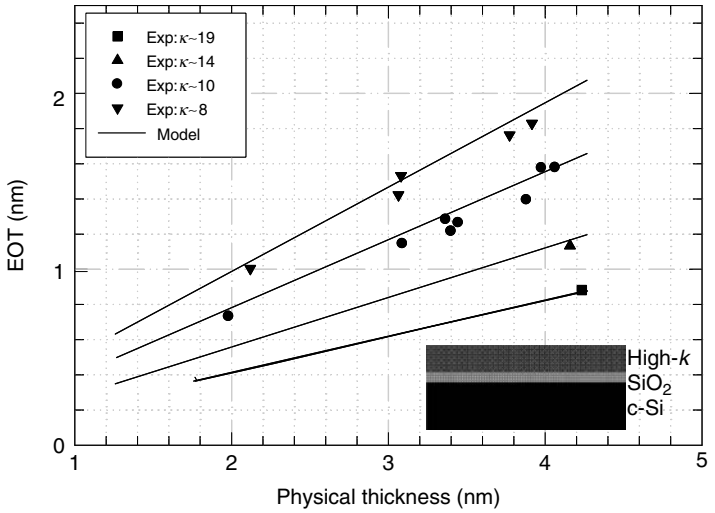


FIGURE 9.15 Calculated effective oxide thickness (EOT) of high-*k* gate dielectric with varying dielectric constant as a function of thickness for a dielectric stack with a constant Si–O thickness of one monolayer (open symbols). The solid symbols represent the data of HfSiON.

9.6.2.4 Electrical Properties

High-*k* gate dielectrics have demonstrated to have much lower gate leakage than SiO_2 and SiON for the same EOT. At this time, HfSiON materials have been fabricated with properties such that they do not significantly degrade the channel mobility for electron and holes. Also according to Sekine et al. [180], plasma nitridation tends to yield higher mobility and better nitrogen profile control in HfSiON. The profile control is important in ensuring that there is minimum interfacial nitridation, which can cause mobility degradation. Mobilities greater than 90% of the universal curve at high electric field for unstrained channels have been achieved [165]. As the introduction of high-*k* gate dielectric is delayed, EOT scaling below 1 nm with high channel mobility is becoming even more critical. Figure 9.16 shows

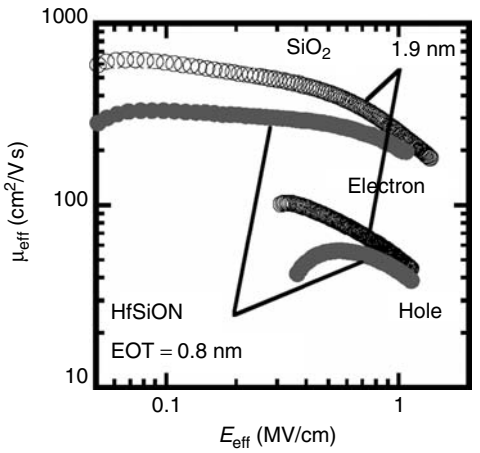


FIGURE 9.16 Effective electron and hole mobility of ultra-thin metal organic chemical vapor deposition (MOCVD) HfSiON and thermal SiO_2 as a function of the effective electric field. (From Inumiya, S. et al., *Extended Abstracts of the 2005 International Conference on Solid State Devices and Materials*, Kobe, Japan, 2005, 10.)

high electron and hole mobilities of HfSiON with an EOT of about 0.8 nm [165]. Hafnium oxide, on the other hand, continues to show lower channel mobility except in cases, where there are indications of a HfSiO interface or an intentional thick SiO_2 interface [181]. A combination of the high channel mobility observed, low charge trapping [182], and ability to scale the EOT suggest that there is hope for scaling of thermally stable high-*k* gate dielectrics.

The principal reason that high-*k* materials have not been introduced using polysilicon gate electrodes yet is in large part because of the pMOS flat band voltage offset. A number of papers have been written on subject, but the precise mechanism for the offset is not yet clearly understood. The offset is most likely due to a combination of factors such as point defects involving dopants in the dielectric and high-*k*/poly silicon interface dipoles [168,183,184]. Because of the large offset and the difficulty in shifting the Vt through channel doping control, the industry is now investigating the use of metal gate electrodes to set the work function and to reduce gate depletion. Most CMOS devices will require at least two metals for threshold control. The use of metal gates also relaxes the equivalent oxide thickness scaling because of the negligible gate electrode depletion. Much work and progress has been made on metal gates and at this point gate electrode depletion has been clearly demonstrated. There are reports that dual work functions can be achieved, however, there are no reports on the use of any of these metals in production yet.

9.7 Summary

9.7.1 Future of SiON as a Gate Dielectric

The 2005 ITRS states for Gate Stack Grand Challenges that (1) extension of oxynitride gate dielectric materials to less than 1.0 nm EOT for high performance MOSFETs, consistent with device reliability requirements and (2) the introduction and process integration of high-*k* gate stack materials and processes for high performance, low operating and low standby power MOSFETs. This raises important questions on how far can SiON be scaled, will a high-*k* dielectric be implemented, how will the thickness and electrical characteristics be measured, and how will the film be manufactured? Rapid thermal processing may offer significant advantages over batch furnace processing, since oxides can be grown at higher temperatures for shorter times and the ambient is easier to control. Surface preparation and passivation are also critical for ultra-thin films.

When considering power dissipation, it appears that gate current leakage of the order of approximately 1 A/cm^2 may be acceptable for some device applications. This gate leakage requirement places a limit on the ultimate scaling limit of about 1.3–1.5 nm on SiON based on recent advances of high nitrogen containing SiON [185,186]. High performance devices that can tolerate higher gate leakage may use SiON thicknesses that are closer to 1 nm. Nevertheless, a high-*k* gate dielectric will be required as we scale EOT lower that 1 nm for high performance devices and lower than 1.3–1.5 nm for low power devices.

9.7.2 Options for SiON Gate Dielectric Replacement

Beyond the 1–1.2 nm regime, a new high-dielectric constant will be required to reduce the gate leakage current. Silicon nitride has proven to extend SiO_2 to approximately 1.2 nm, but even with SiON demonstrated to an EOT of approximately 0.7 nm and acceptable trap density [185,186], gate leakage is still too high for portable low power devices. A thin (1–2 monolayers) layer of SiO_2 has been shown to be required to reduce the interface state density. However, any SiO_2 added to the gate stack will reduce the effectiveness of the higher dielectric constant of silicon nitride. The reason for this is that when adding SiO_2, the nitride thickness must be reduced to keep the same capacitance.

To scale the 0.7–1.0 nm equivalent oxide thickness range, a dielectric with a medium dielectric constant of ≥ 10, an interface oxide < 0.6 nm, and metal gate electrode to eliminate poly silicon depletion will probably suffice. Hf-based oxides (e.g., HfO_2 and HfSiON) have received industry acceptance as the most probable candidate to succeed pure SiON below approximately 1.0 nm.

However, metal oxides and the accompanying gate electrodes may introduce a host of new problems including reduced thermal budgets, potential metal diffusion into the substrate, Fermi Level pinning, etc. The bandgap and barrier heights are important parameters that must also be considered with high dielectric constant dielectrics. Non-ideal currents due to trap-to-trap (Frenkel–Poole) tunneling may also limit the use of high dielectric constant dielectrics.

References

1. ITRS, 2005, in http://www.itrs.net/ITWG/word_files.html
2. Candelier, P. et al. *IEEE Electron Device Lett.* 18 (1997): 306–8.
3. Ragnarsson, L. A. et al. *J. Electrochem. Soc.* 144 (1997): 1866–9.
4. Schay, P. et al. In *Comparison of Gate Oxide Processing Techniques for Thin Dielectric Films*, 41–6. Boston, MA: Materials Research Society, 1996.
5. Berger, M., O. Schwartsglass, and J. Shappir. *J. Electrochem. Soc.* 141 (1994): 2140–5.
6. Devine, R. A. B. et al. In *Near Interface Oxide Degradation in High Temperature Annealed Si/Sio Sub 2 Si Structures*, 623–9. Pittsburgh, PA: Materials Research Society, 1994.
7. Chin, Y. C., M. J. Jeng, and J. G. Hwu. *IEEE Trans. Electron Devices* 45 (1998): 247–53.
8. King, J. C., and C. Hu. *IEEE Electron Device Lett.* 15 (1994): 475–6.
9. Katashiro, M., K. Matsumoto, and R. Ohta. *J. Electrochem. Soc.* 143 (1996): 3771–7.
10. Chao, S. L. et al. *Low Temperature (850°C) Two-Step N sub 2O Annealed Thin Gate Oxides*, 405–40. San Francisco, CA: Materials Research Society, 1996.
11. Yang, T. C., and C. S. Krishna. *Effect of Growth Conditions on the Reliability of Ultrathin MOS Gate Oxides*, 355–60. Pittsburgh, PA: Materials Research Society, 1996.
12. Ajuria, S. A. et al. *IEEE Electron Device Lett.* 17 (1966): 282–4.
13. Misra, V. et al. *IEEE Trans. Electron Devices* 43 (1996): 636–46.
14. Roy, P. K. et al. In *Impact of Various In Situ Preoxidation Process Perturbations on Gate Oxide Quality*, 25–30. San Francisco, CA: Materials Research Society, 1994.
15. Beck, S. E. et al. In *Effects of Microcontaminants in Oxygen During Gate Oxide Growth: Interfacial Effects and Device Reliability*, 100–6. Cambridge, MA: IEEE, 1994.
16. Lin, F. et al. In *Front-End Integration Effect on the Gate Oxide Quality*, 361–6. Pittsburgh, PA: Materials Research Society, 1996.
17. Ligenza, J. R. *J. Electrochem. Soc.* 109 (1962): 73–6.
18. Jorgensen, P. J. *J. Chem. Phys.* 37 (1962): 874.
19. Atalla, M. M. In *Properites of Elemental and Compound Semiconductors*, Vol. 5, edited by H. Gatos, 163–81. New York: Interscience, 1960.
20. Deal, B. E. *J. Electrochem. Soc.* 125 (1978): 576.
21. Flint, P. S. In *The Rates of Oxidation of Silicon*. Los Angeles, CA: Electrochemical Society, 1962.
22. Katz, L. E. In *VLSI Technology*, 2nd ed., edited by S. M. Sze, 98–138. New York: McGraw-Hill, 1988.
23. Massoud, H. Z., J. D. Plummer, and E. A. Irene. *J. Electrochem. Soc.* 132 (1985): 2693–700.
24. Han, C. J., and C. R. Helms. *J. Electrochem. Soc.* 134 (1987): 1297–302.
25. Tiller, W. A., and C. Kang. *J. Cryst. Growth* 2 (1968): 345–55.
26. Massoud, H. Z., and J. D. Plummer. *J. Appl. Phys.* 62 (1987): 3416–23.
27. Gusev, E. P. et al. *Phys. Rev. B* 52 (1995): 1759–75.
28. Massoud, H. Z., J. D. Plummer, and E. A. Irene. *J. Electrochem. Soc.* 132 (1985): 2685–92.
29. Warren, W. L. et al. *Appl. Phys. Lett.* 68 (1996): 2993–5.
30. Tsubouchi, N. et al. *IEEE Trans. Electron Devices* ED-26 (1979): 618–22.
31. Razouk, R. R., L. N. Lie, and B. E. Deal. *J. Electrochem. Soc.* 128 (1981): 2214–20.
32. Lie, L. N., R. R. Razouk, and B. E. Deal. *J. Electrochem. Soc.* 129 (1982): 2828–34.
33. Ahn, S. T. J. et al. *Appl. Phys.* (1989): 65.
34. Xie, J. Z., H. Kauget, and S. P. Murarka. *J. Vac. Sci. Technol.* (1989): B7.
35. Agarwal, A. M., and S. T. Dunham. *J. Electrochem. Soc.* 140 (1993): 222–6.
36. Moharir, S. S., J. Vasi, and A. N. Chandorkar. *J. Inst. Eng. (India)* 76 (1995): 2.

37. Gulbranson, E. A., K. F. Andrew, and F. A. Brassart. *J. Electrochem. Soc.* 113 (1966): 834.
38. Wagner, C. *J. Appl. Phys.* 29 (1958): 1295.
39. Rubloff, G. W. *J. Vac. Sci. Technol.* A8 (1990): 1857.
40. Offenberg, M., M. Liehr, and G. W. Rubloff. *J. Vac. Sci. Technol.* A9 (1991): 1058.
41. Sasse, H. E., and U. König. *J. Appl. Phys.* 67 (1990): 6194.
42. Boukezzata, M. et al. *Thin Solid Films* 279 (1996): 145.
43. Petot, E. G., and C. Petot. *J. Phys. Chem. Solids* 51 (1990): 901.
44. Suzuki, K., and Y. Kataoka. *J. Electrochem. Soc.* 138 (1991): 1794.
45. Jeanjean, P. et al. *Semicond. Sci. Technol.* (1991): 1130.
46. Bureknov, A. F. et al. *Soviet Microelectron. (English Translation of Mikroelektronika)* 17 (1989): 144.
47. Lever, R. F. et al. *Proc. Electrochem. Soc.* 91 (1991): 321.
48. Ishii, H. et al. *Jpn. J. Appl. Phys., Part 2: Lett.* 35 (1996): 1385.
49. Fujise, T. et al. In *Determination of COP Distribution after SC1 Cleaning by a Laser Particle Counter.* San Jose, CA: SPIE, 1996.
50. Abe, T., and Y. Kato. *Jpn. J. Appl. Phys*. 32 (1993): 1879–83.
51. Okada, Y. et al. *IEEE Trans. Electron Devices* 41 (1994): 1608.
52. Fukuda, H., T. Arakawa, and S. Ohno. *Jpn. J. Appl. Phys.* 29 (1990): 2333.
53. Fukuda, H. et al. *IEEE Electron Device Lett.* 12 (1991): 587.
54. Hori, T., H. Iwasaki, and K. Tsuji. *IEEE Trans. Electron Devices* 36 (1989): 340.
55. Momose, H. S. et al. *IEEE Trans. Electron Devices* 41 (1994): 546.
56. Liu, Z. et al. *IEEE Electron Device Lett.* 13 (1992): 519.
57. Joshi, A. B., J. Ahn, and D. L. Kwong. *IEEE Electron Device Lett.* 14 (1993): 560.
58. Dunn, G. J., and P. W. Wyatt. *IEEE Trans. Nucl. Sci.* 36 (1989): 2161.
59. Hori, T., and H. Iwasaki. *J. Appl. Phys.* 65 (1989): 629.
60. Moslehi, M. M., and K. C. Saraswat. *J. IEEE Solid State Circuits* 20 (1985): 26–43.
61. Baumvol, I. J. R. et al. *J. Electrochem. Soc. 143* (1996): 2946–52.
62. Naiman, M. L. et al. *J. Appl. Phys.* 58 (1985): 779.
63. Brow, R. K., and C. G. Pantano. *J. Am. Ceram. Soc.* 70 (1987): 9.
64. Carr, E. C., K. A. Ellis, and R. A. Buhrman. *Appl. Phys. Lett.* 66 (1995): 1492.
65. Okada, Y. et al. *J. Eletrochem. Soc.* 140 (1993): L87–9.
66. Ganem, J.-J. et al. *J. Appl. Phys. Lett.* 68 (1996): 2366–8.
67. Hussey, R. J. *J. Electrochem. Soc.* 143 (1996): 221–8.
68. Bouvet, D. et al. *J. Appl. Phys.* 79 (1996): 7114–22.
69. Gonon, N. et al. *J. Appl. Phys.* 76 (1994): 5242–48.
70. Krüger, D., R. Kurps, and G. Weidner. *Semicond. Sci. Technol.* 8 (1993): 1706.
71. Ting, W. et al. *Appl. Phys. Lett.* 57 (1990): 2808.
72. Okada, Y. et al. *J. Electrochem. Soc.* 141 (1994): 3500.
73. Chu, T. Y. et al. *J. Electrochem. Soc. 138* (1991): L13.
74. Kamath, A. et al. *Appl. Phys. Lett.* 70 (1997): 63–5.
75. Koyama, N. et al. *J. Appl. Phys.* 79 (1996): 1464–7.
76. Kim, K. et al. *J. Electrochem. Soc.* 143 (1996): 3372–6.
77. Peng, J. P., D. Chidambarrao, and G. R. Srinivasan. *Proc. Electrochem. Soc.* 91 (1991): 772.
78. Peng, J. P., D. Chidambarrao, and G. R. Srinivasan. *COMPEL* 10 (1991): 34.
79. Merz, W., and N. Strecker. *Math. Methods Appl. Sci.* 17 (1994): 1165.
80. Kim, K. et al. *Semicond. Sci. Technol.* 11 (1996): 1059.
81. Ling, S.-M., L. H. Dupas, and K. M. D. Meyer. *J. Electron Mater.* 21 (1992): 523.
82. Needs, J. J., K. Board, and C. Taylor. *Microelectron. J.* 22 (1991): 33.
83. Panjukhin, A. V., N. A. Kolobov, and N. A. Leontjeva. *Intl J. Numer. Model.: Electronic Netw. Device Fields* 7 (1994): 15.
84. Toerres, V. J. B. et al. *Interface Sci.* 3 (1995): 133.
85. Schott, K., A. Seidl, and J. Lorenz. *Proc. Electrochem. Soc.* 90 (1990): 417.
86. Umimoto, H., and S. Odanaka. *IEEE Trans. Electron Devices* 38 (1991): 505.

87. Dimitijev, S., H. B. Harrison, and D. Sweatman. In *Model for the Growth of Oxides in N_2O*, 271. San Jose, CA: SPIE, 1994.
88. Dimitrijev, S., H. B. Harrison, and D. Sweatman. *IEEE Trans. Electron Devices* 43 (1996): 267.
89. Sheldon, B. W. *J. Am. Ceram. Soc.* 79 (1996): 2993.
90. Ogata, T. et al. *Jpn. J. Appl. Phys.* 35 (1996): 1690.
91. Pigeon, R. G., and A. Varma. *J. Mater. Sci.* 28 (1993): 2999.
92. Peev, G., L. Zambov, and I. Nedev. *Thin Solid Films* 190 (1990): 341.
93. Pliskin, W. A., and E. E. Conrad. *IBM J. Res. Devices* 8 (1964): 43–51.
94. Archer, R. J. *J. Electrochem. Soc.* 104 (1957): 619.
95. Azzam, R. M. A., and N. M. Bashara. *Ellipsometry and Polarized Light*. Amsterdam, North Holland: Wiley Interscience, 1977.
96. Reisinger, H., H. Oppolzer, and W. Honlein. *Solid State Electron.* 35 (1992): 797–803.
97. Snyder, P. G. et al. *J. Appl. Phys.* 60 (1986): 3292.
98. Chao, T. S., C. L. Lee, and T. F. Lei. *J. Electrochem. Soc.* (1991): 138.
99. Pliskin, W. A., and R. P. Resch. *J. Appl. Phys.* 36 (1965): 2011–3.
100. Corl, E. A., and H. Wimpfheimer. *Solid State Electron.* (1964): 7.
101. Reizman, F., and W. V. Gelder. *Solid State Electron.* 10 (1967): 625.
102. Nicollian, E. H., and J. R. Brews. *Metal Oxide Semiconductor (MOS) Physics and Technology*. New York, NY: Wiley-Interscience, 1982.
103. Spence, J. C. H. *Experimental High-Resolution Electron Microscopy*. 2nd ed. Oxford: Oxford University Press, 1988.
104. Cherns, D. *Analytical Techniques for Thin Film Analysis*. San Diego, CA: Academic Press, 1988.
105. Buseck, P. R., J. M. Cowley, and L. Eyring, eds. *High-Resolution Transmission Electron Microscopy and Associated Techniques*. New York: Oxford University Press, 1988.
106. Sheng, T. T. In *Analytical Techniques for Thin Film Analysis*, edited by K. N. Tu, and R. Rosenberg, 252–96. Boston, MA: Academic Press, 1988.
107. Tu, K. N., and R. Rosenberg, eds. *Analytical Techniques for Thin Film Analysis*. Boston: Academic Press, 1988.
108. Runyan, W. R., and T. J. Shaffner. *Semiconductor Measurements and Instrumentation*. New York: McGraw-Hill, 1998.
109. Kar, S., and W. E. Dahlke. *Solid State Electron.* 15 (1972): 221–37.
110. Wu, E. et al. *Semicond. Sci. Technol.* 15 (2000): 425–35.
111. Depas, M. et al. *Solid State Electron.* 37 (1994): 433–41.
112. Lehovec, K., and S. T. Lin. *Solid State Electron.* 19 (1976): 993–6.
113. Zafar, S. et al. *Appl. Phys. Lett.* 67 (1995): 1031–3.
114. Hauser, J. R., and K. A. Ahmed. In *Characterization and Metrology for ULSI Technology*, edited by D. G. Seiler, 235–8. Woodbury, NY: AIP, 1998.
115. Wolf, S. *Silicon Processing for the VLSI Era—The Submicron MOSFET*. Vol. 3. Sunset Beach, CA: Lattice Press, 1995.
116. Wolters, D. R., and J. J. V. D. Schoot. *Philips J. Res.* 40 (1985): 115.
117. Schuegraf, K. F., and C. Hu. *IEEE Trans. Electron Devices* 41 (1994): 761–7.
118. DiMaria, D. J., and J. W. Stasiak. *J. Appl. Phys.* 65 (1989): 2342–56.
119. DiStefano, T. H., and M. Shatzkes. *Appl. Phys. Lett.* 25 (1974): 685–7.
120. Chen, I.-C., S. E. Holland, and C. Hu. *Solid State Circuits IEEE J.* 20 (1985): 333–42.
121. Harari, E. *J. Appl. Phys.* 49 (1978): 2478–89.
122. Ricco, B. et al. *Phys. Rev. Lett.* 51 (1983): 1795.
123. Nissan-Cohen, Y., and T. Gorczyca. *IEEE Electron Device Lett.* 9 (1988): 287.
124. Lee, J. C., I.-C. Chen, and C. Hu. *IEEE Trans. Electron Devices* ED-35 (1988): 2268.
125. Sze, S. M. *Physics of Semiconductor Devices*. 2nd ed. New York: Wiley-Interscience, 1981.
126. Hu, C. et al. *IEEE Trans. Electron Devices* ED-32 (1985): 375.
127. Heremans, P. et al. *IEEE Trans. Electron Devices* (1988): 35.
128. Deal, B. E. *IEEE Trans. Electron Devices* ED-27 (1980): 606–8.

129. Schroeder, K. *Semiconductor Material and Device Characterization*. New York: Wiley Interscience, 1990.
130. Iwata, S., and A. Ishizaka. *J. Appl. Phys.* 79 (1996): 6653–709.
131. Yang, H. et al. *Electron Device Lett.* 21 (2000): 76.
132. Koike, K., K. I. S. Nakamura, G. Inoue, A. Kurokawa, and S. Ichimura. *J. Electron. Mater.* 34 (2005): 240.
133. Momose, H. S. et al. *IEEE Trans. Electron Devices* 43 (1996): 1233–42.
134. Rana, F., S. Tiwari, and D. A. Buchanan. *Appl. Phys. Lett* 69 (1996): 1104–6.
135. Hattangady, S. V. et al. In *1996 International Electron Devices Meeting,* 495-8. San Francisco, CA: IEEE, 1996.
136. Niimi, H., and G. Lucovsky. *J. Vac. Sci. Technol.* B17 (1999): 2610.
137. Yang, H., and G. Lucovsky. In *1999 International Electron Device Meeting Technical Digest*, 245. Washington, DC: IEEE, 1999.
138. Ellis, K. A., and R. A. Buhrman. *J. Electrochem. Soc.* 145 (1998): 2068.
139. Hattangady, S. V., H. Niimi, and G. Lucovsky. *Appl. Phys. Lett.* 66 (1995): 3495.
140. Grider, D. T. et al. In *Symposium on VLSI Technology Digest of Technical Papers*, 47. Kyoto, Japan: The Japan Society of Applied Physics, 1997.
141. Khamankar, R. et al. In *Symposium on VLSI Technology Digest of Technical Papers*, 162. Honolulu, HI: The Japan Society of Applied Physics, 2004.
142. Inaba, S. et al. In *2002 International Electron Device Meeting Technical Digest*, 651. San Francisco, CA: IEEE, 2002.
143. Chen, C.-C. In *Symposium on VLSI Technology Digest of Technical Papers*, 176. Honolulu, HI: The Japan Society of Applied Physics, 2004.
144. Niimi, H. et al. *J. Appl. Phys.* 91 (2002): 48.
145. Nicollian, P. E. et al. In *2000 International Electron Device Meeting Technical Digest*, 545. San Francisco, CA: IEEE, 2000.
146. Kawae, T. et al. In *International Workshop on Gate Insulators*, 146. Tokyo, Japan: The Japan Society of Applied Physics, 2003.
147. Young, J. T., P. M. Lenahan, and G. J. Dunn. *IEEE Trans. Nucl. Sci.* 39 (1992): 2211.
148. Liu, C. H. et al. In *Reliability Physics Symposium Proceedings,* 268. Dallas, TX: The Electron Device Society and the Reliability Society of the IEEE, 2002.
149. Lucovsky, G. et al. *J. Vac. Sci. Technol. B* 17 (1999): 1806–12.
150. Niimi, H., and G. Lucovsky. *J. Vac. Sci. Technol. A* 17 (1999): 3185.
151. Lucovsky, G., and J. Phillips. *J. Non-Cryst. Solids* 227 (1998): 1221.
152. Yugami, J. et al. In *International Workshop on Gate Insulators,* 140-5. Tokyo, Japan, 2003.
153. Wang, Y. R. et al. In *Symposium on VLSI Technology Digest of Technical Papers,* 164. Kyoto, Japan: The Japan Society of Applied Physics, 2005.
154. Robertson, J. *J. Vac. Sci. Technol. B* 18 (2000): 1785–91.
155. Visokay, M. R. et al. *Appl. Phys. Lett.* 80 (2002): 3183.
156. Shanware, A. et al. In *2001 Technical Digest of IEEE International Electron Device Meeting*, 137-40. Washington, DC: IEEE, 2001.
157. Rotondaro, A. L. P. et al. *IEEE Electron Device Lett.* 23 (2002): 603–5.
158. Ino, T. et al. In *Extended Abstracts of the 2005 International Conference on Solid State Devices and Materials*, 230-1. Kobe, Japan, 2005.
159. Koike, M. et al. In *2003 Technical Digest of IEEE International Electron Device Meeting*, 4.7.1-4. Washington, DC: IEEE, 2003.
160. Guha, S. et al. *Appl. Phys. Lett.* 81 (2002): 2956–8.
161. Fischetti, M. V., D. A. Neumayer, and E. A. Cartier. *J. Appl. Phys.* 90 (2001): 4587–608.
162. Kerber, A. et al. In *Reliability Physics Symposium Proceedings, 2003. 41st Annual. 2003 EEE International,* 41-5. Dallas, TX: The Electron Device Society and the Reliability Society of the IEEE, 2003.
163. Rotondaro, A. L. P. et al. *IEEE Electron Device Lett.* 23 (2002): 603–5.

164. Rotondaro, A. L. P. et al. In *Symposium on VLSI Technology Digest of Technical Papers,* 148-9. Honolulu, HI, 2002.
165. Inumiya, S. et al. In *2003 Symposium on VLSI Technology Digest of Technical Papers*, Kyoto, Japan: The Japan Society of Applied Physics, 2003.
166. Rotondaro, A. L. P. et al. In *Symposium on VLSI Technology Digest of Technical Papers,* 148-9. Honolulu, HI, 2002.
167. Hobbs, C. et al. In *2003 Symposium on VLSI Technology Digest of Technical Papers*, Kyoto, Japan: The Japan Society of Applied Physics, 2003.
168. Kamimuta, Y. et al. In *2005 International Conference on Solid State Devices and Materials*, 24-5. Kobe: The Japan Society of Applied Physics, 2005.
169. Koyama, M. et al. *Technical Digest of IEEE International Electron Device Meeting* (2004): 752.
170. Zhang, Z. B. et al. *Symposium on VLSI Technology Digest of Technical Papers* (2005): 50–1.
171. Wilk, G. D., R. M. Wallace, and J. M. Anthony. *J. Appl. Phys.* 87 (2000): 484–92.
172. Shanware, A. et al. In *2001 Technical Digest of IEEE International Electron Device Meeting*, 137-40. Washington, DC: IEEE, 2001.
173. Ritala, M., and M. Leskelä. In *Handbook of Thin Film Materials*, Vol. 1, edited by H. S. Nalwa, 103–59. San Diego, CA: Academic Press, 2002.
174. Senzaki, Y. et al. *J. Vac. Sci. Technol. A: Vac. Surf. Films* 22 (2004): 1175–81.
175. Hendrix, B. C. et al. *Appl. Phys. Lett.* 80 (2002): 2362–4.
176. Inumiya, S. et al. In *2003 Symposium on VLSI Technology Digest of Technical Papers*, 17-8. Kyoto, Japan: The Japan Society of Applied Physics, 2003.
177. Visokay, M. R. et al. *Appl. Phys. Lett.* 80 (2002): 3183–5.
178. Yamaguchi, T. et al. *IEEE International Reliability Physics Symposium* (2003): 34–40.
179. Morais, J. et al. *Appl. Phys. Lett.* 86 (2005): 212903–6.
180. Sekine, K. et al. In *2003 Technical Digest of the International Electron Devices Meeting*, 4.6.1-4. Washington, DC: IEEE, 2003.
181. Callegari, A. et al. In *2004 Technical Digest of the International Electron Device Meeting*, San Francisco, CA: IEEE, 2004.
182. Shanware, A. et al. In *2003 Technical Digest of the International Electron Devices Meeting*, 38.6.1-4. Washington, DC: The Japan Society of Applied Physics, 2003.
183. Hobbs, C. C. et al. *Electron Device IEEE Trans.* 51 (2004): 971–7.
184. Kaneko, A. et al. In *Extended Abstracts of the 2003 International Conference on Solid State Devices and Materials*, 56-7. Tokyo, Japan: The Japan Society of Applied Physics, 2003.
185. Tsujikawa, S. et al. In *2002 Symposium on VLSI Technology Technical Digest,* 202-3. Honolulu, HI, 2002.
186. Matsushita, D. et al. In *Technical Digest of IEEE International Electron Device Meeting*. Washington, DC. 2005.

10

Silicides

10.1 Scope of the Chapter **10**-1

10.2 Introduction **10**-2
Brief Historical Account of Contacts for Integrated Circuits • The Self-Aligned Silicide (Salicide) Process

10.3 Development Trends of Salicide **10**-5
Ti Silicide Process-Challenges and Solutions • Co Silicide Process—Challenges and Solutions

10.4 Nickel Silicide for Contacts and Interconnections **10**-18
Basic Properties of Ni Silicide Phases • Reactive Phase Formation • Formation Mechanism • Challenges and Progress • Anomalous Thermal Expansion of NiSi and Related Stress Effects • Texture Development in NiSi Films • $Si_{1-x}Ge_x$ Devices

10.5 Metal Gate and Schottky Barrier Source-Drain **10**-38
Perspectives • Fully Silicided Gates: Dual Effective Work-Function • Silicide Schottky Barrier Source-Drain

10.6 Summary **10**-43

Acknowledgments **10**-43

References **10**-44

Christian Lavoie
Francois M. d'Heurle
IBM Thomas J. Watson Research Center

Shi-Li Zhang
Royal Institute of Technology and Fudan University

10.1 Scope of the Chapter

The goals of this chapter are to give an overview of silicide contacts to microelectronic devices and to present many important material advantages and limitations for microelectronic applications. A simple web search on the current subject easily leads to more than 10,000 references, without counting the numerous books and precursor work not yet in electronic format. This chapter is therefore not an exhaustive review but a summary point of view of the authors, suitable for the current book.

Metal silicides have played an important role in improving circuit performance as the miniaturization of the device dimensions continues in pace with the predictions of Moore's law. The primary use of silicides evolved from creating reliable contacts for simple diodes, to generating high conductivity paths for local wiring, and lately to forming low-resistivity electrical contacts for metal-oxide-semiconductor field effect transistor (MOSFETs) using a self-aligned silicide process. The present chapter gives a short historical account of contacts to devices, followed by a description of the self-aligned process, and of the suitable characteristics of compatible silicides. The development trends are then presented in which we cover the main limitations of the self-aligned silicides widely used in the industry, namely Ti and Co silicides. As it becomes clear that Ni silicide will be the material for contacts to state-of-the art logic devices, an extensive section covers this material. The chapter then touches on fully silicided gates (FUSI)

and the possibility of utilizing alternative and complementary silicides to reduce Schottky barrier heights in order to limit contact resistance in the most advanced devices.

10.2 Introduction

10.2.1 Brief Historical Account of Contacts for Integrated Circuits

Over the past 40 years, the methods used to form the contacts to semiconductor devices in integrated circuits have first evolved from direct metal deposition and subsequent patterning to the exploitation of metal-silicon compounds. The first reported use of silicide material in electrical devices dates to the early 1960s, where PtSi was used as contact between Al and Si to improve the rectifying characteristics of diodes [1]. Early in the 1970s, PtSi [2] and Pd_2Si [3,4] were used as the contact layer between Al and Si, which ensured a good surface coverage and reduced the resistance of the contact. PtSi was also important in bipolar technologies where a Schottky diode was incorporated between the base and the collector of the device to form a "clamped transistor," [5] effectively reducing the saturation time by limiting minority carrier transport. In the 1980s, the silicides appeared in large-scale integration of MOSFETs and were initially used to reduce the resistance of poly-Si interconnections. In the original MOSFET integrations, the local interconnections between devices were indeed made of doped poly-Si with a resistivity of the order of 500 $\mu\Omega$ cm. As line width of these interconnections reached below 1 μm, the resistance became unacceptable and prompted a search for less resistive materials. The presence of the poly-Si was necessary to keep the properties of the interface SiO_2/poly-Si intact and because the processing is in relatively oxidizing atmosphere, conducting silicides became very attractive. The first silicide used, WSi_2, reduced the resistivity of the interconnections by an order of magnitude to 50 $\mu\Omega$ cm [6]. The original stack thus included a poly-Si layer directly in contact with the gate oxide, a layer containing both Si and metal and a capping SiO_2 layer. The layers were then thermally processed to form the required silicide compound and patterned to define the gate structures and local interconnections [6]. This process, named polycide, was limited to silicides that could withstand the high temperatures of dopant activation. A restricting limitation inherent to the process is that reactive ion etching of the silicide must be possible. In other words, the metal selected must form halide compounds with high vapor pressures. These criteria mainly limited the selected metals to Ti, Ta, Mo, and W [7].

The polycide process in which the formation of silicided gates occurs early in the device fabrication, was followed in the early 1990s in large scale integration by a self-aligned process in which the silicides are formed simultaneously over the poly-Si interconnections as well as the source, drain, and gates of field effect transistors after devices are already functional [8–11]. This process will be the main focus of this chapter. Introduced more than 15 years ago, the self-aligned silicide contacts have already evolved multiple times as their formation has been successively based on Ti, Co, and Ni. These process modifications throughout the years lead to significant and necessary improvements in performance or robustness. Each of the metallizations was tailored for a given generation of microelectronic devices and was thus consistent with the process flow at the time. Although often regarded as separate entities, the unit processes evolve simultaneously and any modifications to one lead to constraints on another. The whole process often changes sufficiently so that many new elements with new and optimized properties would be practically incompatible with prior processes. The recent use of NiSi represents a good example of such an evolution. This most advanced process for forming silicide contacts would surely be incompatible with any of the process flows used for prior generations of integrated circuits.

Early metallization was based on direct Al contacts. A description of the evolution of these contacts was given in [12] and will be summarized here. The Al was deposited directly on the Si and then patterned using lithography to leave the Al only in desired areas. A schematic of the resulting structure for contact to a single p–n junction is presented in Figure 10.1a. The patterned Al is directly in contact with the Si. As the junction depth in the Si became shallower because of the continued scaling, the electrical properties of devices using such contacts were found to be much affected by the additional, subsequent thermal treatments after contact fabrication. As the dimensions were reduced, the devices were also very sensitive

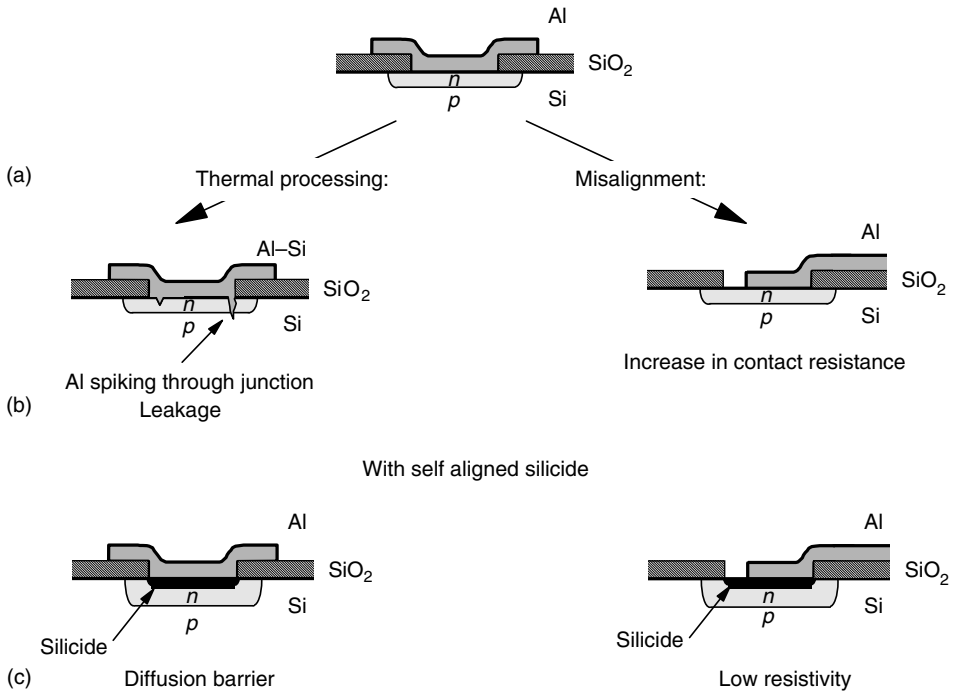


FIGURE 10.1 Schematics describing early metallization. (Adapted from d'Heurle, F. M. *J. Electron. Mater.* 27, 1138, 1998.)

to the alignment of masks defining the contact. First, the post contact anneals were found to lead to Al spiking in the Si, possibly shorting the electrical junction below the contact (Figure 10.1b). The formation of these spikes was caused by the dissolution of Si into the Al. As the temperature increases, the solid solubility of Si in Al increases to more than 1 at% at 500°C and close to 2 at% at the eutectic temperature of 577°C [13]. The first solution to avoid dissolution of Si in Al was to deposit an Al–Si alloy already containing Si above the solubility limit. While fixing the dissolution problem, this also led to a slight increase in the resistivity of the material. More importantly, since the solid solubility of Al in Si decreases with decreasing temperature, upon cooling the devices, some Si precipitates epitaxially at the Si–Al interface. This Si is p-doped as it is saturated in Al and thus gives rise to an increased contact resistance especially on *n*-type Si substrate. As the size of the contacts reach the size of the epitaxial islands, this process simply becomes unusable. The electrical characteristics of the device were also very dependent on the alignment of the contacts. As also shown in Figure 10.1b, any misalignment of the Al did lead to a reduction in contact area, a critical factor for the injection of the carriers in the device.

Figure 10.1c shows how the addition of a silicide layer directly formed on the source, drain and gate of transistors before metallization and patterning can help to solve both problems above. Indeed, the silicide limits the diffusion of Si and also renders the contact relatively insensitive to misalignments of the Al because of the low resistivity of the selected silicide. As long as the Al contact touches the silicided area, the carriers distribute quickly through the silicide and uniformize the carrier transport even if the contact is not perfectly aligned.

10.2.2 The Self-Aligned Silicide (Salicide) Process

The formation of the silicide over the exposed Si areas of an integrated circuit is performed using a self-aligned process schematically presented in Figure 10.2. The term self-aligned refers in this case to

a process for which resist and lithography steps are not necessary [8–12]. This method dramatically reduces the complexity and cost of forming the first contact to the Si areas. The process flow is described for a standard MOSFET build on a Si on insulator substrate. In Figure 10.2a, after critical surface cleaning steps, the wafer is coated with a thin metal layer often directly followed by the deposition of a capping layer to prevent oxidation during subsequent anneals. While the first industrial process using Ti did not require the use of a capping layer, the particular sensitivity of Co silicide formation to the presence of oxygen required the presence of a capping layer. The most recent Ni silicide formation has been successful in many circumstances without capping layers but its sensitivity to oxide and oxidation, although not as high as for Co silicide formation, is important so cap layers may be advantageous depending on environment and process conditions.

Figure 10.2b shows the transistor after a first anneal to form silicide over the exposed Si areas. The temperature of this anneal cannot be so high as to allow Si diffusion in the metal leading to a silicided connection shorting the source and gate of transistors, a phenomenon referred to as bridging. The annealing process must be carefully selected to form a silicide phase that resists the etch solutions capable of removing metal layers. After this anneal, the metal is then selectively etched typically in solutions of hydrogen peroxide (H_2O_2) and sulfuric acid (H_2SO_4). Once the metal is removed (Figure 10.2c) and with it the possibility of bridging, a second anneal can be performed to obtain a more suitable silicide phase if necessary. This is certainly the case for Ti and Co silicides. For Ni silicide, however, it is feasible to use a single step anneal but many have argued that a formation with two anneals is still desirable [14–18]. Nowadays, after completing the silicide formation, the device wafer is capped with interlayer dielectrics and vias are then opened to make connections to the silicide (10.2d). While these process steps are not specifically related to silicide formation and will be covered in a different chapter of this book, they are mentioned here to emphasize the importance of the silicide as an etch step for the opening of the via

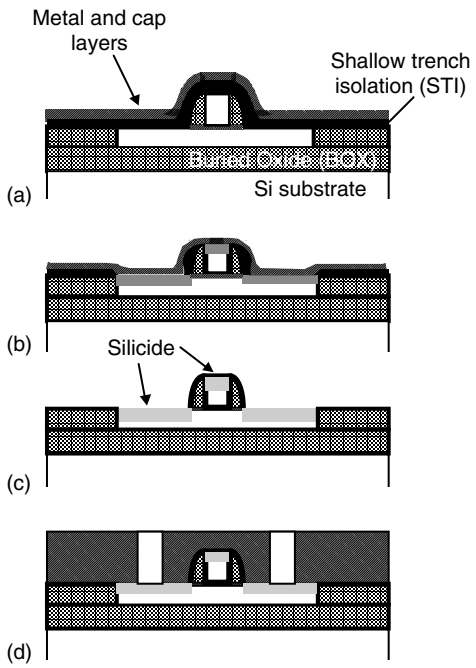


FIGURE 10.2 Schematics describing the self-aligned silicide process.

holes. The silicide must resist the reactive ion etching so that a contact can be formed between the back end metallization and the device. The etch chemistry must be adapted so that no volatile compounds can be formed with the silicide. While the formation of halides is clearly documented, other compounds such as volatile carbonyls used in the chemical vapor deposition of these metals may significantly affect the etching.

From the schematics in Figure 10.1 and Figure 10.2, we can deduce the following desired properties for a silicide material.

Low resistivity (limits contact alignment issues and reduce device resistance)
Etch selectivity of the silicide vs. the metal (allows self-aligned process)
Etch resistance in reactive ion etch (RIE) environment (allows opening of via holes)
Acceptable diffusion barrier properties
Low roughness (gives a minimal junction penetration)
Preferably high resistance to oxidation.

Beside these six characteristics, the silicide must also meet the following criteria:

High morphological stability (films must retain their shapes throughout the back-end-of-line (BEOL) processing)
Minimal Si consumption (limited doped Si available)
Controlled film stress.

With the device dimensions decreasing and the junctions becoming shallower with each technology generation, one must realize that these criteria are constantly evolving. For example, simple criteria for limited roughness or high morphological stability were clearly less restrictive 10 years ago than they are today. In particular, the requirement to form an acceptable diffusion barrier for Si is relatively non-existent nowadays since a diffusion barrier is formed before the next metal level is deposited. The last three criteria above have become more important in order to meet the requirements set by the fabrication of current, very small devices. First, it becomes necessary to limit the processing temperatures after the high temperature anneal (commonly called "activation anneal") necessary to force solutions of the dopants within the Si lattice. Any subsequent anneal may precipitate the dopants and reduce the conductivity of the junction. Limiting the allowable thermal budget thus helps to decrease the effect of "dopant deactivation." This leads to the use of materials forming at much lower temperatures and therefore also materials that tend to degrade at lower temperatures. As an example, efforts have been recently focused on stabilizing NiSi against agglomeration. Limiting Si consumption is also more critical since the amount of Si available for silicide formation becomes limited in thin silicon-on-insulator (SOI) or shallow junctions. Finally, current device optimization commonly includes layers and process steps to engineer the stress in the device channel in order to enhance carrier mobility. In such situations, it is necessary that the silicide formation do not disturb the pre-engineered stress conditions. Furthermore, it is desired that stress levels in the silicide be uniform across a wafer and consistent for all formed dimensions.

While multiple silicides could represent decent compromises with respect to the criteria mentioned above, the industry has converged towards the three lowest resistivity silicides for the self-aligned process, namely $TiSi_2$, $CoSi_2$, and NiSi, which are all in the range of 10–25 $\mu\Omega$ cm. With silicides of thickness reaching from about 20 to 30 nm, the resistivity range depends on film thickness, grain structure and impurity concentration within the layer.

10.3 Development Trends of Salicide

In this section, a more comprehensive description of the $TiSi_2$ and $CoSi_2$ processes with their advantages and challenges is given, following an overview of the trends in the materials properties, as the silicide evolves first from C54-$TiSi_2$ to $CoSi_2$ and then to NiSi. These trends can be followed in Table 10.1. While

TABLE 10.1 Comparisons of Silicide Properties for C54-$TiSi_2$, $CoSi_2$, and NiSi

Silicide Property/Characteristic	C54-$TiSi_2$	$CoSi_2$	NiSi
Thin film resistivity (μΩ-cm)	15–25	15–20	10–20
Formation temperature (°C)	750–850	600–750	300–500
Melting temperature (°C)	1500	1326	992
Si consumed normalized to metal thickness/silicide thickness	2.2/0.91	3.6/1.03	1.8/0.83
Silicide thickness normalized to metal thickness/Si consumption	2.4/1.1	3.5/0.97	2.2/1.20
Total volume contraction in percentage metal thickness (final interface position)	C54-$TiSi_2$ → 78%	CoSi → 82%	Ni_2Si → 42%
		$CoSi_2$ → 110%	NiSi → 80%
Controlling formation mechanism	Nucleation	Nucleation/Diffusion	Diffusion
Diffusing species	Si	Co, Si, Co[a]	Ni
Schottky barrier height to *n*-Si (Ev)	0.60	0.64	0.67
Limitations	Transformation C49 → C54	Rise in R_s of narrow lines, roughness, SiGe	High temperature degradatation

[a] Co for Co_2Si, Si for CoSi and Co for $CoSi_2$ if formed at low *T*. For $CoSi_2$ at high temperature, both elements diffuse similarly.

there are some slight advantages for thin film resistivity and Si consumption, the main advantages reside in the change of formation mechanisms and reduction in formation temperatures. Moving from nucleation controlled to diffusion controlled formation, leads to silicide layers that are smoother and to reactions that are much more uniform and predictable. When a reaction is nucleation controlled, the thermodynamic drive for the reaction to occur is low as the change in free energy from the reactant to the product is small. As a result, the number of nucleation sites is small, the film typically rough and the formation becomes highly dependent on a multitude of process parameters (dopants, substrate types, device dimensions, anneal conditions, etc.). In the case of a diffusion controlled formation, as will be discussed later, the change in Gibbs free energy is large, the reaction is much less dependent on process parameters, and, the films are typically much smoother. Compared to $TiSi_2$ and $CoSi_2$ films, the NiSi of similar thicknesses typically show an order of magnitude reduction in roughness. The reduction in formation temperature is also a critical advantage. As mentioned earlier, high temperatures tend to deactivate some of the dopants (push dopants that are within the Si lattice into precipitates), which lead to an increase in the depletion length in the poly-Si just above the gate oxide. Dopant deactivation also leads to an increase in contact resistance between the silicide and the Si. This contact resistance is exponentially dependent on the Schottky barrier height and the activated dopant concentration. While the barrier heights are relatively close for the three materials, the high temperatures necessary for the Ti and Co silicide processes are detrimental to device performance and NiSi shows advantages such as a decreased dopant deactivation.

Although the trends observed typically lead to improved performance, it is important to notice that they are not necessarily the motivation for a change. As the devices are scaled to ever smaller dimensions, the current material faces important limitations. A new material is of course selected to be more performing but the change almost always results from the limitations of the earlier material. Faced with continuous scaling, the optimized material is plagued with problems that may not be alleviated easily. This represents the point at which the industry is willing to invest large amounts of money to implement new processes and new materials.

Some difficulties of a material set facing this continuous scaling are described in the next two subsections covering successively Ti and Co silicide processes.

10.3.1 Ti Silicide Process-Challenges and Solutions

The Ti silicide process was the first one to be widely implemented in the so-called salicide technology for complimentary metal oxide semiconductor (CMOS) devices. Compared to the possible alternatives that lead to low resistivity silicides, the Ti process is much less sensitive to Si oxidation at the surface. The oxidation level depends, for example, on varying dopant implantation conditions. As a result of its capability to reduce native oxide and dissolve the oxygen, Ti was very well suited for generations or devices where the cleaning of surfaces was not so demanding as what can be industrially achieved today. Titanium disilicide may exist in two different crystallographic structures: the C49-$TiSi_2$ phase, usually considered to be metastable and the stable C54-$TiSi_2$ phase. The resistivity of the C49 structure is typically four times that of the C54 (70–100 μΩ-cm vs. 15–25 μΩ-cm). Although the C54-$TiSi_2$ is thermodynamically favored, undetermined factors lead first to the nucleation and formation of the high resistivity C49-$TiSi_2$ phase. It is not clear why the C49 phase is favored but it could depend on different surface energies, on slightly different diffusion coefficients in the two phases or on stress relaxation. The C49 is indeed prone to form stacking faults since the nearly tetragonal structure allows for an easy crystallographic rotation of the building blocks. The growth of the C49 structure usually begins at about 500°C–600°C followed by the transformation to the C54 phase above 700°C. The phase sequence and variation in material characteristics is best described using rapid in situ monitoring techniques during annealing of thin films. The anneal of a 32 nm Ti film deposited on standard poly-Si is presented in Figure 10.3. The characterization apparatus combines x-ray diffraction (XRD), elastic light scattering and resistivity measurements, enabling the simultaneous measurements of changes in phases, texture, resistivity, and surface roughness as reactions between metal films and silicon substrate proceed. These experiments were performed at the NSLS X20C beamline of Brookhaven National Laboratory (U.S.A.). The XRD energy is selected to be 6.9 keV (wavelength = 0.18 nm) using a multilayer monochromator with a 1.5% energy resolution that provides an x-ray flux greater than 10^{13} photons/s. This intensity corresponds to a gain of four orders of magnitude over a standard rotating anode system. The combination of the high x-ray photon flux with a linear position sensitive detector allows (a) fast

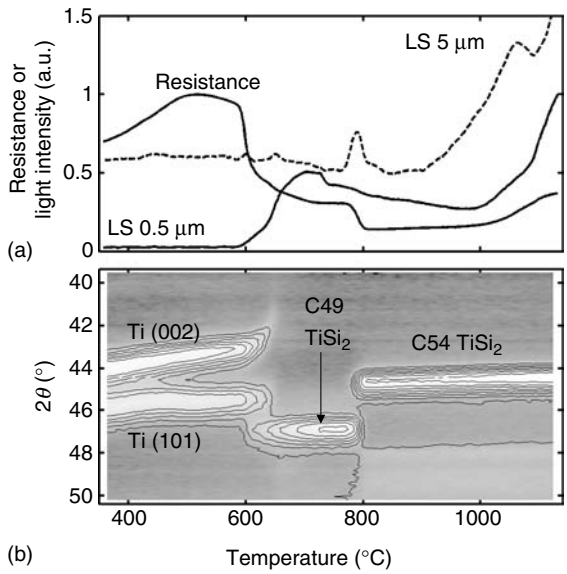


FIGURE 10.3 In situ measurements of during annealing at 3°C/s of a 32 nm Ti film deposited on a poly-Si substrate (a) resistance and light scattering (0.5 and 5 μm) signals, (b) simultaneous x-ray diffracted intensity. (Adapted from Lavoie, C., Cabral, C., d'Heurle, F. M., and Harper, J. M. E., *Defect Diffus. Forum*, 1477, 194–199, 2001.)

data acquisition during rapid thermal anneals (RTA), (b) observation of very thin films and other things being equal, (c) considerably improved signal to noise ratios. In the current configuration, a diffraction spectrum (10–15 in 2θ) from a 10 nm metal film can be acquired in less than 100 ms. The measurements of roughness are made from laser light with a wavelength of 633 nm, at an incident angle of 65° and scattered angles of 20 and 52° (with two detectors) providing information on length scales of about 0.5 and 5 μm, respectively. The electrical resistance is measured using a spring loaded four point probe in a square geometry. Anneals presented in this chapter are performed in purified He and at a heating rate of 3°C/s. Temperature measurements are calibrated using metal-silicon eutectics and are precise to ± 3°C. Further experimental details are provided elsewhere [19–22].

In Figure 10.3b, the x-ray intensity is presented both as a gray scale (black to white from low to high intensity) and as intensity contours. For temperatures up to 600°C, two x-ray peaks are observed at $2\theta \sim 43.7°$ and $2\theta \sim 45.6°$ that correspond to the diffraction from (002) and (101) planes of crystalline hcp Ti. For this temperature range, the two scattering signals A and B are essentially featureless. The resistance, however, increases linearly up to about 500°C and quickly drops above 600°C as the Ti x-ray peaks decrease. Simultaneously, the small length-scale scattering signal starts increasing as the C49-$TiSi_2$ phase forms.

As the temperature is increased to 750°C, we observe a stabilization of the resistance, the 0.5 μm scattering signal and the diffracted intensity of the (131) planes of the C49-$TiSi_2$ phase. For slightly higher temperatures, the silicide film clearly transforms into the low resistivity C54 phase, as seen from the x-ray data and resistance trace. The (311) peak of the C54-$TiSi_2$ appears slightly below 800°C and persists to over 1100°C, the end of the temperature ramp. At the C54 formation, the long length scale scattering intensity goes through a sharp maximum, a signature often observed for nucleation controlled transformation. Since the reaction proceeds from isolated nucleation centers, the middle of the transformation correspond to a mix of large patches of C49 and C54-$TiSi_2$, which are being detected by the 5 μm length-scale detector. Once the transformation is completed, the C54 film returns to pre-transformation levels of roughness.

Above 900°C, both scattering signal and resistance trace show important variations related to the high temperature degradation of the film (while the XRD seems insensitive to it). The peak at about 1060°C for the 5 μm light scattering is due to silicide-poly-Si inversion [23], while the continuous roughening that occurs is a consequence of agglomeration.

In the self-aligned process described earlier, the first anneal is around 700°C and brings the film into the C49 phase. At such a temperature Si diffusion is already relatively high and early attempts to form the silicide lead to Si diffusion along the grain boundary of the Ti and to electrical bridges between the gate and source/drain of transistors. Performing the anneals in N_2 was found to be critical to make the process possible. Nitrogen diffuses in the grain boundaries of Ti and limits the diffusion of Si.

The transition from the C49 to the C54-$TiSi_2$ is difficult because of the small driving force that is the difference in free energies between the two phases (0.48 kJ/g-atom) [24]. This small difference is of the order of the error in measurements performed to obtain the respective values. Nucleation is difficult and the density of nuclei low, so that the line resistance becomes uncontrollable for $TiSi_2$, when formed in pre-defined Si lines below 0.5 μm in width [25,26]. This effect is related to both the line width and the area of Si to be silicided. The phenomenon is clearly observed when comparing the XRD of Figure 10.3 with the two XRD plots of Figure 10.4. This last figure emphasizes the C49–C54 transformation in narrow lines of 0.35 μm in width. These samples are formed of many thousands of identical lines, which are measured simultaneously. Only the transformation is measured as the C49-$TiSi_2$ has already been formed in each of the narrow structure using the standard salicide process. For a blanket film of Ti, the transformation was shown to occur around 800°C at the experimental ramp rate of 3°C/s. In Figure 10.4a, for the same temperature ramp rate, a structure of 0.35 μm in width and 11 μm^2 in area (about 31 μm in length) does transform in the C54 phase only at about 930°C. As the area of the structure is reduced from 11 to 3 μm^2 (from about 31 to less than 9 μm in length), the transformation temperature remains similar but the transformation only proceeds partially. While for the larger areas, the C49 peak

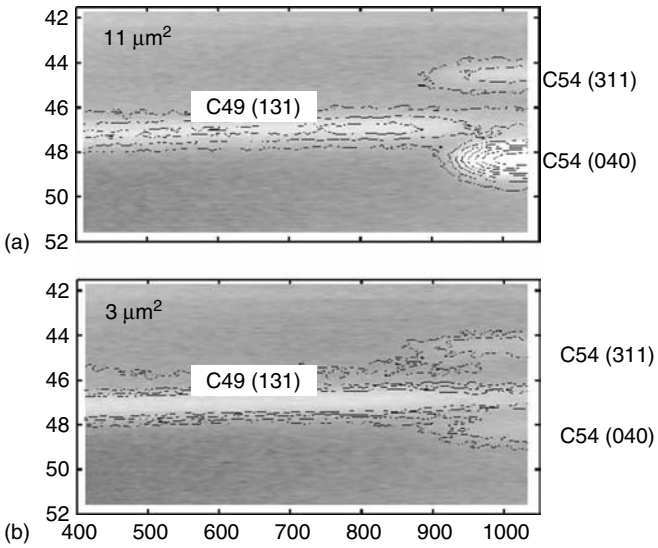


FIGURE 10.4 In situ XRD measurements of C49–C54 transformation in lines of 0.35 μm wide for nominal Ti thickness of 32 nm. The x-ray beam probes more than 10^5 identical rectangular structures of (a) 11 μm^2 and (b) 3 μm^2. For (b) the transformation is clearly incomplete as the C49 peak remains to the highest temperatures.

did completely disappear upon C54 formation, the same peak now only decreases slightly, clearly showing an incomplete transformation.

The phenomenon observed is obviously related to the low nucleation density and is best described using the schematic of Figure 10.5. In (a) the transformation in blanket films is first depicted. In the area selected, only three nucleation centers for C54 are present among the many small C49 grains. As the transformation proceeds at 800°C and each nucleus is allowed to grow, the three nucleation centers grow into three large C54 grains up to complete consumption of the C49 phase. In (b) the same area is divided into four lines. From Figure 10.4a, the transformation occurs considerably higher at 930°C and only happens in the lines containing the most active nucleation sites. This rise in temperature is associated with a variation of the texture of the C54. In the blanket films, most C54 grains have the (311) orientation mostly parallel to the surface while in narrow lines the (040) orientation dominates. As the length of the line is reduced, in (c), the transformation is clearly only partial but its temperature remains the same and is found to be related to the line width.

We are therefore facing two phenomena, a clear one of nucleation density, which leads to incomplete transformation and a second one of difficult C54 growth along narrow lines, which leads to an increase in transformation temperature. The nucleation density for C54 can roughly be evaluated from the grain size. C54 grains typically have a diameter of about 2 to more than 3 μm defining a typical nucleation density of about 0.1–0.3 μm^{-2}. This density should be highly dependent on the preparation conditions leading to the C49. We have observed C54 grains to reach a diameter of 60 μm in some films suggesting that the nucleation density can be lower by almost three orders of magnitude, which would lead to the transformation problem observed at much larger dimensions. For the transformation to be complete, at least one nucleation center must be present in each of the defined small areas. The difficult growth along narrow lines is not so well understood and much harder to study independently since as the temperature increases one expects other nucleation centers to become active. It is however clear that the transformation in the narrow lines shown in Figure 10.4 is not initiated at 800°C, the temperature of the transformation in blanket films. We are therefore dealing with a transformation controlled either by

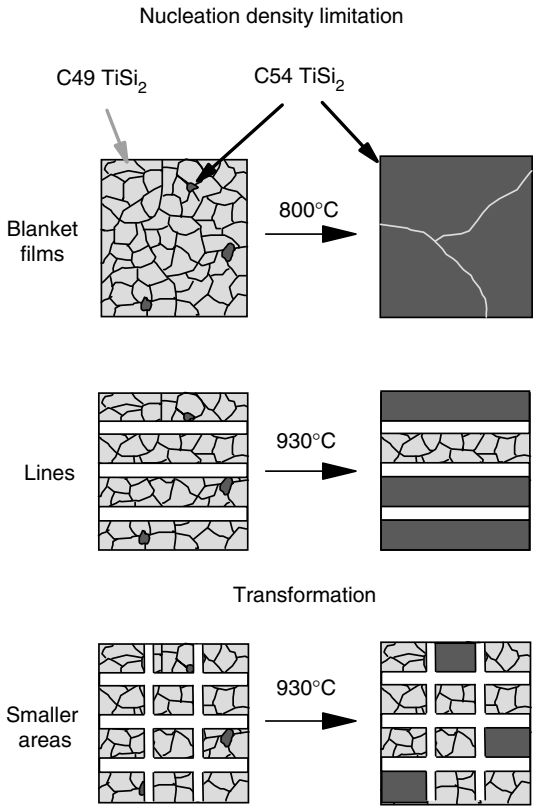


FIGURE 10.5 Schematics explaining the lack of C49–C54 transformation in small dimensions based on the limited number of nucleation centers.

different nucleation centers, by the difficult growth along narrow dimensions, or by a combination of both effects. Hypothesis related to texture and growth velocities have been presented in the past [27,28].

With the origin of the difficulty in the C54-$TiSi_2$ formation at least partly identified, various solutions have been proposed in order to increase the nucleation density of C54. Noteworthy among numerous attempts, the three following solutions are directly applicable in manufacturing. The first solution is the utilization of pre-amorphization implantation (PAI) prior to Ti deposition and silicidation [29–31]. The motivation behind PAI was to create a C49-$TiSi_2$ structure with more defects and smaller grains leading to an increased number of grain junctions and boundaries where the C54-$TiSi_2$ grains are believed to be preferentially nucleating. Enhanced formation of C54-$TiSi_2$ in narrow Si lines was demonstrated using Si, Ge, B or As implants in which C49 grain size was reduced at least by a factor of 3 reaching only 70 nm in diameter. This process was applied successfully in line width close to 100 nm.

The second developed solution is the addition of small percentages of refractory metals (RM) such as Mo, Ta, Nb, and W to the Ti–Si system. Intensive research on the addition of RM began after an accidental breakthrough at IBM for which ion implantation of Mo and W was found to decrease the formation temperature of C54-$TiSi_2$ by 100°C–150°C and to eliminate the bimodal distribution in resistance observed for narrow line width structures. The combination of Mo implantation with PAI has lead to formation of C54-$TiSi_2$ in lines of only 60 nm in width [29]. First experiments introducing RM in

the Ti–Si system has been performed by means of implantation [32,33], of thin interlayer of Mo, Ta or Nb [34–36] or by means of Ti-RM alloy deposition [37]. The transformation behaviors to C54-$TiSi_2$ for a pure Ti film and for a Ti(5.5 at% Ta) are presented in Figure 10.6. The Ti image is the same as in Figure 10.4b. For the same structures, the addition of a few atomic percents of Ta does not only reduce the C54 formation temperature but also much increases the fraction of C54 formed. While any of the way of adding RM to the system produces similar effects, the addition of the RM directly to the sputtering target is the most elegant solution as it only requires a change in sputtering material and does not require any additional processing steps as for implantation or deposition of extra interlayers. Multiple explanations for the enhancement have been proposed. The first one, as in the case of PAI, is a reduction in C49 grain sizes [38]. Others such as the template effect suggest that the grains of C40-(Ti-RM)Si_2 form first and serve as template for the nucleation of C54-$TiSi_2$. The C11b tetragonal structure of high temperature $MoSi_2$ and WSi_2, the C40 hexagonal structure of $CrSi_2$, $NbSi_2$, $TaSi_2$, and low temperature $MoSi_2$ and WSi_2, as well as the C54 orthorhombic $TiSi_2$ all share the same hexagonal lattice arrangement of the basal plane: one metal atom surrounded by 6 Si atoms. These three structures differ only by the stacking sequence of these planes with an ABA stacking for C11b, ABCA stacking for C40 and ABCDA stacking for C54. If the C40 or C11b phase can form, it is reasonable to assume that the C54 could nucleate on the very similar basal plane. An increase in the electron-to-atom ratio has also been suggested as a reason for stabilizing the C54 structure as seen for example on the transition from equilibrium C54-$TiSi_2$ to equilibrium C49 in Ti(Si, Al)$_2$ [39]. The variation of the number of electrons changes the position of the Fermi level within the density of state function, which does not vary much within these proposed structures. Structural calculations show a stabilization of the C54 phase consistent with observations.

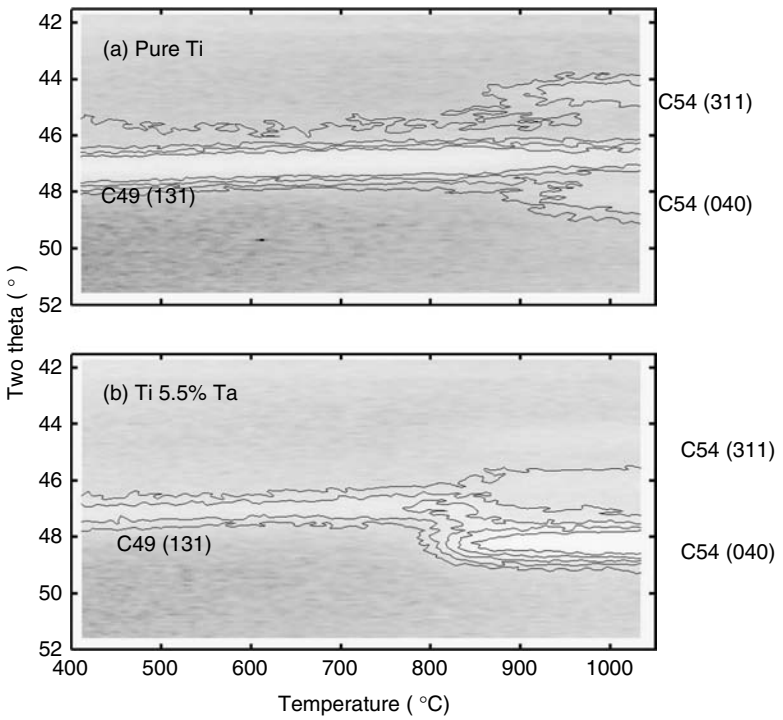


FIGURE 10.6 In situ XRD measurements of C49–C54 transformation temperature in lines of 0.35 μm wide and area of 3 μm^2. The x-ray beam probes more than 10^5 identical rectangular structures built from metal films of (a) 32 nm Ti and (b) 32 nm Ti (5.5 at% Ta). The addition of Ta clearly decreases the transformation temperature.

The true mechanisms responsible for the observed enhancement effects remain to be confirmed and debates still continue. Most likely, the complexity of the system ensures that several factors are at play. For details about the RM addition to enhance the formation of C54, a review may be consulted [40].

The third proposed solution is to control the first anneal process. Multiple or long anneals performed on the C49 before the final transformation anneal at higher temperature were shown to enhance the C54 formation. Annealing excessively for many hours significantly decreased the apparent activation energy for the C49–C54 transformation from the typically high value of 4 to 6 eV to a lower value of 3 eV [41]. Excimer laser annealing was shown to enhance the formation of C54, by first imposing the formation of the C40 phase also unstable like the C49 phase. Upon a second anneal, after the standard selective etch, the formation of C54 was observed to form at temperatures closer to 600°C [42–44]. The mechanism for this is not clear at this point. It is particularly surprising that the C54 would be easier to form from a C40 matrix as the nucleation barrier is expected to be even higher from this phase; the C40 is structurally much closer to the C54 than the C49 one. While the mechanisms are not clearly understood, it is apparent that variations in anneal treatment can cause drastic changes in the formation of the low resistivity C54 phase. Since laser annealing techniques are becoming available as they are being used in manufacturing for rapid dopant activation, this alternative may be worth revisiting for $TiSi_2$ metallization on state-of-the-art devices.

10.3.2 Co Silicide Process—Challenges and Solutions

Although some of the modifications to the $TiSi_2$ process allowed for relatively inexpensive solutions for dimensions down to about 0.25 μm, further decreases in gate lengths required the introduction of $CoSi_2$. This low resistivity material did not show formation problems in the smallest achievable dimensions at that time ($\sim$100 nm).

The Co silicide formation sequence is presented in Figure 10.7 (3°C/s). At the temperatures of interest for Si processing, the equilibrium diagram for the Co–Si system shows only three phases: Co_2Si, CoSi, and $CoSi_2$. With the reaction of Co thin films with Si substrates, these three phases are formed sequentially in that order [45]. As for the reaction of Ti and Si shown earlier, the reaction of an 8 nm Co film with an undoped SOI (100) substrate is also presented using measurements of scattered light and resistance (Figure 10.7a) as well as XRD (Figure 10.7b). The XRD peak at about 52° is that of hexagonal Co (002). The change, which is observable above 350°C is a consequence of the abnormal growth of grains with the (001) orientation [46] and corresponds to a slight decrease in the resistance rather than the expected monotonic increase with temperature as the phonon population increases also. At higher temperature (450°C), the increase in XRD intensity at a slightly larger angle than for the Co (002) peak corresponds to the appearance of the Co_2Si (301) peak. The formation of Co_2Si leads to a steep increase in resistance, which stabilizes around 500°C when the monosilicide phase forms as evidenced by the appearance of the (210) and (211) XRD peaks of CoSi at about 53 and 58°C. These peaks are much less intense as the texture of the monosilicide phase is highly random compared to the other phases present. The peak just above 55°C present at temperatures higher than 600°C is the (220) diffraction line of $CoSi_2$. For this uncapped Co film, the two scattered light signals, also presented in Figure 10.2a show drastically different behaviors. Both detected signals are relatively featureless until the formation of the disilicide. At that temperature, the signal from lateral length scales of about 0.5 μm increases monotonically up to about 700°C and is then stable for about 150°C before a second increase starting at about 850°C which is due to morphological degradation (agglomeration). The roughening observed on the 0.5 μm length-scale is typical of $TiSi_2$, $CoSi_2$, and $NiSi_2$ formation where the nucleation barrier is important and most of the roughness develops on lateral length-scales larger than 200 nm. As for the C54-$TiSi_2$ formation shown earlier, the long length-scale detector (5 μm) only shows a very sharp peak resulting from large lateral non-uniformity during formation. At the mid point of the formation, areas of CoSi and $CoSi_2$ coexist on these length scales generating a temporary roughness that disappears as the transformation reaches completion [47]. The lateral non-uniformity in the $CoSi_2$ formation emphasizes that nucleation

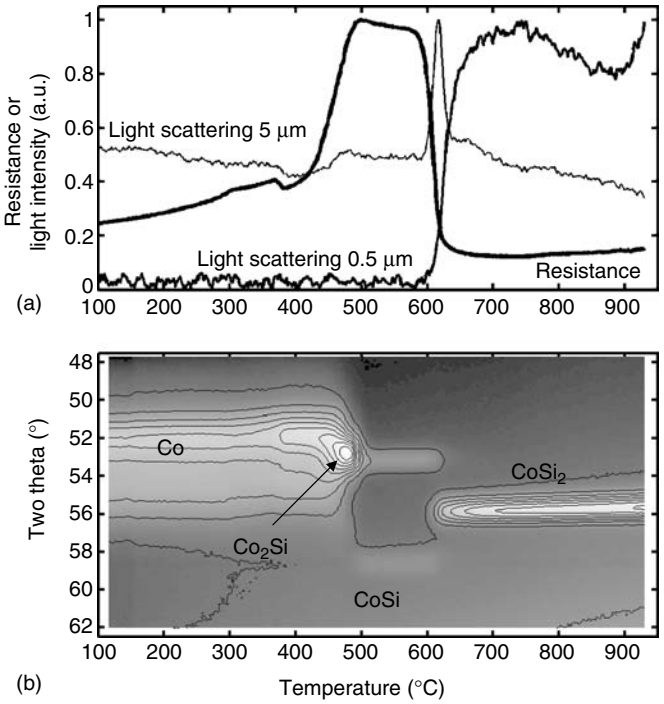


FIGURE 10.7 (a) Resistance and light scattering from 0.5 to 5 μm length scales together with (b) x-ray diffraction measurements performed in situ during annealing (3°C/s) of a 9-nm thick Co film deposited on a Si on insulator (SOI) substrate. (Adapted from Lavoie, C., Cabral, C., d'Heurle, F. M., and Harper, J. M. E., *Defect Diffus. Forum*, 1477, 194–199, 2001.)

is a relatively important factor for this material set. As a result of this importance of nucleation, the $CoSi_2$ film is inherently rough.

In the self-aligned process described in Figure 10.2, the first anneal step for the Co process is usually around 450°C–580°C and forms the smooth monosilicide phase. After the selective etch, a second anneal typically in the range of 650°C–750°C forms the rougher $CoSi_2$ phase with the desired low resistivity. An anneal at the higher temperature of 750°C (as opposed to 650°C) does not seem to be necessary from Figure 10.7 in which the Si is undoped. However, with high doses of dopants, the nucleation temperature of $CoSi_2$ can easily increase by 100°C. In particular, the nucleation is sensitive to high doses of As. We also studied the effect of alloying elements on the formation and morphology of CoSi and $CoSi_2$ films [48]. Of the 23 alloying elements investigated, many showed a smoother film than the unalloyed $CoSi_2$ but most significantly increased the temperature of its nucleation. Only Cu and Ni additions lead to a significant decrease in the disilicide formation temperature. Large increases in resistivity were however observed with the Cu additions.

When gate lengths reach dimensions significantly shorter than 50 nm, multiple factors limit the formation of the low resistivity phase $CoSi_2$. There are four main factors that limit the continued use of this material in future devices:

The sensitivity to contamination from oxygen, both at the surface and in the annealing atmosphere,
The limited volume of Si available for the reaction as the junctions become very shallow and the SOI becomes very thin,
The introduction of Si–Ge,
The rise in resistance in very narrow lines,

10.3.2.1 Cleaning of Surfaces and Oxidation

One difficulty in implementing a Co silicide process resided in the optimization of the cleaning process before Co deposition. While Ti has the ability to reduce any surface oxide upon annealing, a Co process is much more sensitive to the presence of oxygen either in the annealing atmosphere or directly at the Si surface. Figure 10.8 describes the sensitivity of the Co process to the degree of oxidation of the surface before Co deposition. With increasing oxide thickness, the observed behavior of phase formation is not straightforward. Processes that completely remove the oxide from a Si surface and prevent oxidation before Co deposition lead to a very rough surface with facets and spikes extending through the junction. The silicide spikes have been observed [49] to follow (111) Si planes deep into the substrate. We have seen these spikes reaching ten times deeper (>200 nm) than the thickness of the mono silicide (~20 nm). With a very thin suboxide, the resulting $CoSi_2$ shown in the second cross section of Figure 10.8 contains this oxygen rich layer (bright) within the film and the layer becomes relatively smooth and adequate for integrated circuits. An increase in oxide thickness leads to direct epitaxial formation of $CoSi_2$, a phenomenon referred to as oxide mediated epitaxy (OME) [50,51]. An oxide that is sufficiently thick is presumed to act as a diffusion barrier and to slow the rate of arrival of the Co atoms to allow direct epitaxy. This type of epitaxial formation was found to result in encroachment under the sidewall spacers (as seen in the plan-view Transmission emission microscope (TEM) image of Figure 10.8) and to facets at the interface. The variation in oxide thickness described here is extremely small, as a native oxide found on Si areas is sufficient to completely block the reaction between Co and Si.

The sensitivity to oxide at the interface is also accompanied with an extreme sensitivity to the oxygen present in the atmosphere during the first anneal. This required the Co process to always include a cap to prevent oxygen penetration during the first anneal. TiN and Ti caps have both been used extensively in the industry. The use of a reactive Ti cap was shown to reduce the effect of contamination from surrounding areas and effectively eliminate the thinning of $CoSi_2$ observed at the edge of a contact [52]. The TiN cap (unreactive) results in an increased process window but the surface cleaning processes

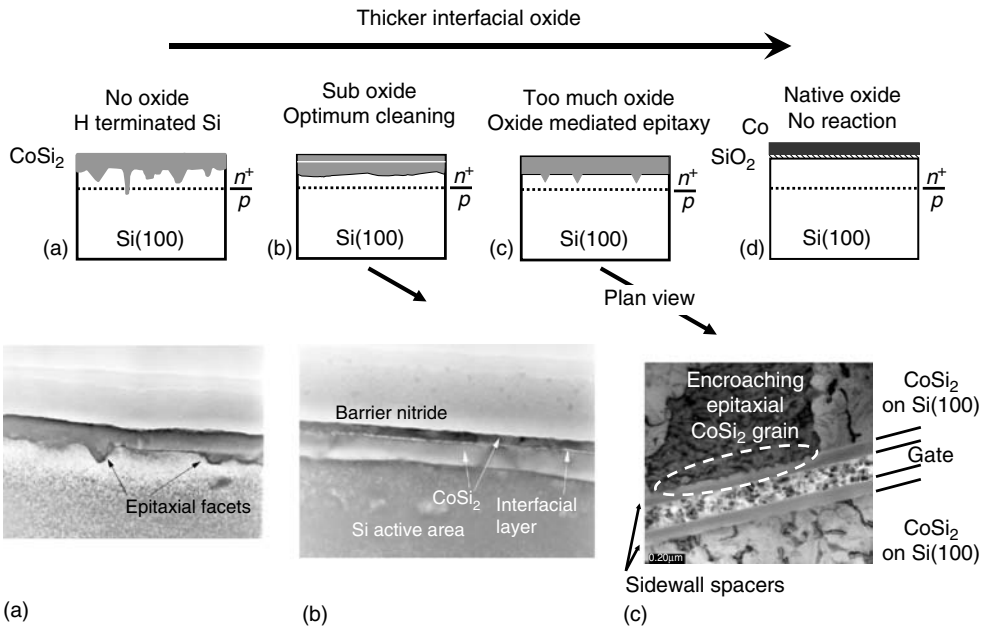


FIGURE 10.8 Schematics of the effects of various surface cleaning on $CoSi_2$ roughness combined with descriptive scanning electron microscopy images.

become more critical. In particular, the combination of TiN caps with sputter-cleaning before metal deposition can be problematic because of the redistribution of impurities from the oxide and nitride structures.

The importance of the cap (TiN or Ti) over the Co film as a barrier to oxygen contamination during the first anneal can be understood from Figure 10.9 where uncapped Co films, 8 nm thick, deposited on a Si substrate were annealed in an atmosphere of He with different impurity contents. When using our standard purification system (Figure 10.9a), the oxygen concentration level is much lower than the part per billion level ($\ll 10^{-9}$). In this case, the phase sequence observed is the same as if a TiN cap is used. For Figure 10.9b, the pumping time was reduced (little desorption of H_2O) and the purifier was turned off and the He (99.9995%) flowed directly from the gas cylinder, with the oxygen content in the "ultra high purity" He of the order of one part per million ($\sim 10^{-6}$). In this case, the in situ XRD clearly shows the extension of the Co (002) peak to temperatures up to the point ($\sim$600°C) where an extra peak appears at about 52° (2θ). We believe this peak to be related to a compound of Co, Si, O, and possibly N. While the formation of CoSi was found to be very sensitive to the oxygen content of the annealing atmosphere, no significant differences were observed during the formation of $CoSi_2$ from CoSi in the same conditions. A capping layer is not necessary during the second anneal of the salicide process.

10.3.2.2 Reduction in Available Si for the Reaction

The much shallower junction depths and the use of thin SOI also present serious challenges for scaling down of the $CoSi_2$ process. To meet the sheet resistance requirements of current technologies, the thickness of the $CoSi_2$ film must be ranging from 20 to 30 nm. A simple calculation comparing the crystal structures and dimensions of Si and $CoSi_2$ shows that the formation of the low resistivity silicide requires a Si layer that is 3% thicker than the silicide itself [53]. While this volume change is very small, it is still interesting that for a given volume, there is more Si in $CoSi_2$ than in Si, a consequence of a change in the atomic bonding from covalent (longer bond) to more metallic in nature. The Si consumption is then

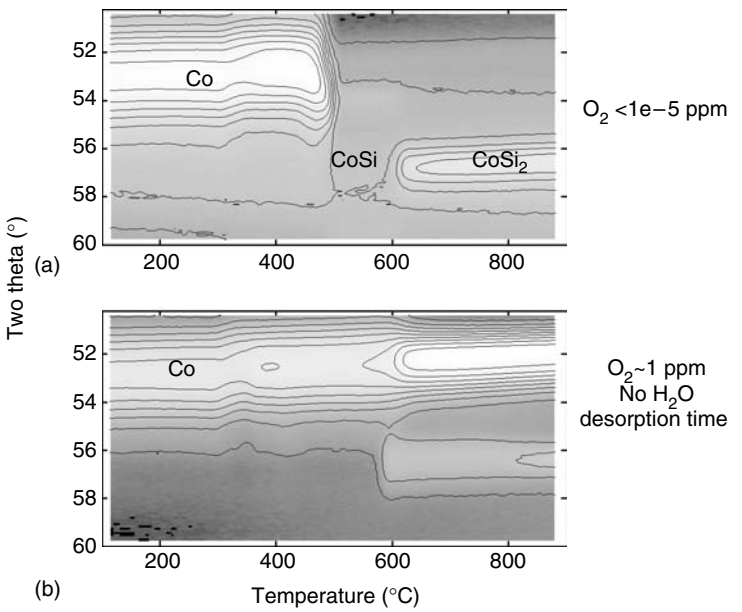


FIGURE 10.9 Effect of oxygen contamination (and presence of H_2O desorption) on the Co silicide formation from a Co film 8 nm thick on Si(100). The O_2 concentration in He was (a) less than 10^9 and (b) about 10^{-6}. For (b) the pumping time was minimal and some water desorption may be present. (3°C/s). (Adapted from Lavoie, C., Cabral, C., d'Heurle, F. M., and Harper, J. M. E., *Defect Diffus. Forum*, 1477, 194–199, 2001.)

about the average thickness of the silicide film. However, in terms of leakage through a shallow junction, one must consider the roughness of the film, as the local maximum thickness of the film will be the point of high leakage. The $CoSi_2$/Si interface is inherently rough because of the importance of nucleation in the formation of $CoSi_2$ from CoSi [7,54–57]. Although an optimized cleaning procedure or some low level of alloying can reduce the film roughness significantly [48,58], some interface roughness remains and the maximum layer thickness is typically 20%–30% larger than the average $CoSi_2$ thickness. When the thickness of the SOI layer or the junction depth reaches about 40 nm, at least part of the silicide film will touch the underlying oxide layer or reach the junction causing significant degradation in contact resistance and device properties.

10.3.2.3 Introduction of Ge

Another factor restricting the extended use of $CoSi_2$ is the addition of Ge within the substrate. Germanium is also a semiconducting material but with a smaller gap (0.67 eV vs. 1.10 eV for Si) and exhibiting much higher carrier mobilities [59]. Its atomic volume is slightly larger than that of Si with which it shows complete miscibility [13]. The use of SiGe either as a blanket layer or locally on source and drain of pFET devices allows for a modification of the stress in the device channel thereby increasing carrier mobility and device switching speed. The formation of $CoSi_2$ on such material is extremely difficult [48,60,61]. While CoGe is soluble in CoSi, $CoGe_2$ is apparently immiscible in $CoSi_2$. The CoGe and CoSi both have the same cubic crystal structure (prototype FeSi) while the structures of $CoGe_2$ and $CoSi_2$ are orthorhombic (prototype $PdSn_2$) and cubic respectively (prototype CaF_2). From classical nucleation theory [57], it can be shown that the increased entropy of mixing for the solution Co(Si, Ge) increases the barrier for nucleation [48,57,61,62]. As a result, the addition of 10–20 at% Ge to the Si leads to an increase in the nucleation temperature of $CoSi_2$ from about 600°C to close to or above 800°C and it also leads to a decrease in nucleation density as can be directly observed from optical microscopy [48]. From in situ measurements of phase formation, we have determined that not only the nucleation is elevated to higher temperatures but also the growth of the phase is much retarded after nucleation has occurred. This phenomenon is best observed in Figure 10.10 in which we compare the reaction of an 8 nm Co film with an SOI substrate (a) and a Si capped $Si_{0.8}Ge_{0.2}$ substrate (b). The layer of Si is sufficiently thick (20 nm) to allow complete formation of the monosilicide ($\sim$15 nm needed). Both reactions proceed very similarly up to about 650°C at which temperature the $CoSi_2$ starts to form. For the silicide film forming on the Si substrate (Figure 10.10a), the formation occurs over a narrow temperature range. For the film with the Si/SiGe underlying substrate (Figure 10.10b), the formation is much more gradual and finishes close to 800°C. Since there is Si still available between the CoSi and the SiGe, the nucleation occurs at the same temperature as in Figure 10.10a, but when the growing film reaches the SiGe, the formation is dramatically hindered and although nucleation has clearly occurred, the presence of the Ge delays the formation. Although these temperatures are relatively close to the Ge melting point, the requirement that the Ge be expelled from a growing $CoSi_2$ grain matters since Si and Ge interdiffusion in SiGe is still very slow. This increase in formation temperature is too high for manufacturing of advanced devices as it leads to significant dopant deactivation. More importantly, for SiGe with typical Ge content (20–30 at%), a thin continuous $CoSi_2$ film of low resistance cannot be formed. Indeed, the presence of the low melting point Ge limits the morphological stability of the film as the agglomeration temperature is reduced to the point where a continuous film cannot form before the agglomeration process begins.

10.3.2.4 Rise in Resistance (Voiding) in Narrow Poly-Si Lines

The resistance of $CoSi_2$ lines increases dramatically with decreasing line width with dimensions below about 50 nm [63–68]. Figure 10.11 reproduced from [22] shows the resistance of narrow n and p polysilicon lines as a function of line width for both the $CoSi_2$ and NiSi process. As will be discussed in the next main section, the well behaved NiSi formation in these dimensions is one important reason for current shift towards this technology. The resistance increase observed for the Co process is caused by the presence of voids within the narrowest silicide lines [66]. The width at which the voids start to be present depends on the process, the shape, and the dimensions of the silicided structure (line length and

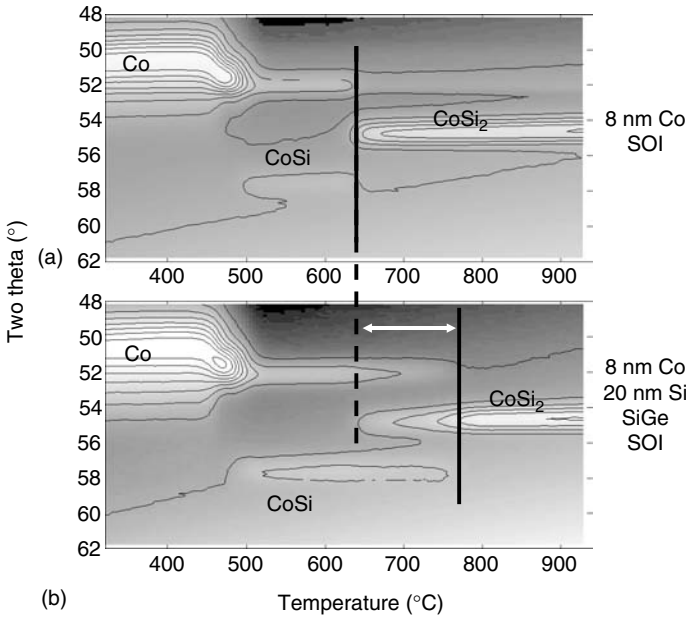


FIGURE 10.10 In situ XRD of the silicide phase formation from a Co film 8 nm thick on (a) Si(100) and (b) Si/SiGe. In (b), the Si is sufficiently thick to allow the formation of CoSi and the nucleation of $CoSi_2$ at similar temperatures as in (a). When the formation of $CoSi_2$ reaches the SiGe layer, the formation is delayed by more than 100°C. (3°C/s). (Adapted from Lavoie, C., Cabral, C., d'Heurle, F. M., and Harper, J. M. E., *Defect Diffus. Forum*, 1477, 194–199, 2001.)

geometry). The origin of these voids is not clear and may depend on the presence of impurities, on surface preparation, on early and non-uniform agglomeration in smaller dimensions, on local stresses, or even on the mechanisms of formation (diffusion or nucleation controlled formation, diffusing species). During the formation of CoSi, Si is the dominant diffusing species. As a result, multiple vacancies tend to precipitate in the underlying poly-Si lines. Even if this precipitation of vacancies in narrow dimensions plays an important role, we are certainly confronted with a combination of the above factors. Attempts to

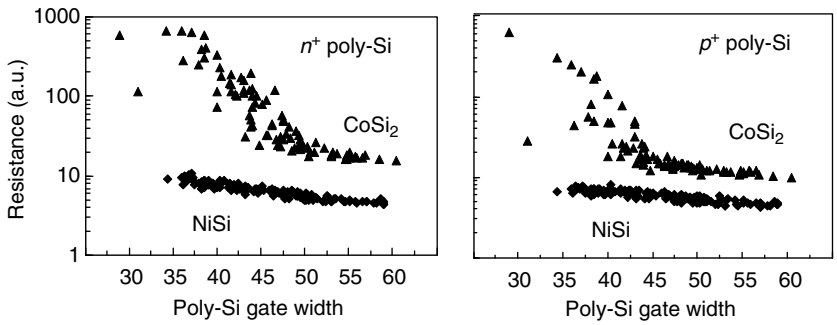


FIGURE 10.11 Resistance of patterned poly-Si lines as a function of line width for $CoSi_2$ and NiSi. The resistance of $CoSi_2$ lines on (a) *n*-doped and (b) *p*-doped poly-Si increases in narrow gate dimensions while NiSi resistance is independent of width. (Reproduced from Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004. With permission.)

alleviate this problem using a Co–Ni alloy which reduces the $CoSi_2$ formation temperature [48,62,69] were not successful [61,70].

Some of the factors limiting the extended use of $CoSi_2$ can be partially alleviated through the optimization of the material by alloying [48]. The selective addition of epitaxial Si to the gate, source and drain of transistors (raised source/drain process; RSD) eliminates most of the concerns and would allow for a continued $CoSi_2$ process. However, besides the increased process complexity and cost, the fairly long anneals at high temperature negatively affect dopant profiles and causes significant dopant deactivation. The possibility of a NiSi process becomes extremely interesting because of the lower temperatures and both reduced cost and process complexity.

10.4 Nickel Silicide for Contacts and Interconnections

The convergence towards Ni based salicide for ohmic contact formation in the source drain and gate regions of a MOSFET is a combined result of the technological evolution and requirements, our understanding of the materials properties and our mastering of the process details. In this section, technologically relevant issues, challenges, likely solutions and practices related to Ni-salicide are presented.

10.4.1 Basic Properties of Ni Silicide Phases

10.4.1.1 Phase Diagram

The Ni–Si binary phase diagram in Figure 10.12 shows a rather complex picture with up to eleven phases, six of which are stable at room temperature (Ni_3Si, $Ni_{31}Si_{12}$ or Ni_5Si_2, Ni_2Si, Ni_3Si_2, NiSi, and $NiSi_2$) [13]. This fact may considerably increase the complexity of phase formation sequence to be treated in Section 10.4.2. It may also complicate the dependence of phase formation on processing parameters and substrate variations (dopant type and concentration, cleaning conditions). An important observation is the relatively low melting point of the desired low-resistivity phase NiSi, 990°C, which is a primary cause for the much concerned morphological instability of NiSi thin films that will be discussed in Section 10.4.3. Furthermore, the presence of the thermodynamically stable $NiSi_2$ has an important implication

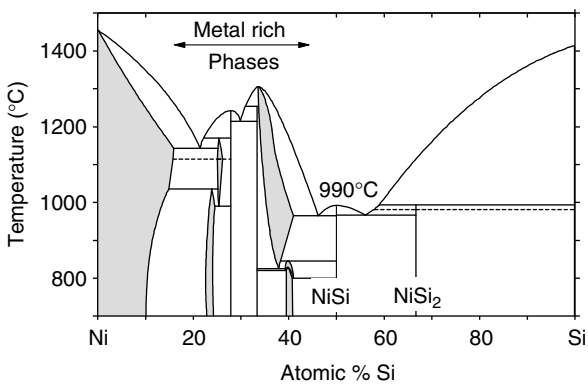


FIGURE 10.12 Binary phase diagram for Ni–Si. Note the complexity in number of stable phases at room temperature and the low temperature for melting. (Adapted from Handbook of Binary Alloy Phase Diagrams (CD version 1.0), ASM international and Reproduced from Lavoie, C., d'Heurle, F. M., Detavernier, C., Cabral, C., *Microelectron. Eng.* 70, 144, 2003 and Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004. With permission.)

for the use of NiSi: in the presence of excess Si, NiSi is unstable and $NiSi_2$ should form. It is interesting to note in the phase diagram that about 10 at% Si can be dissolved in Ni at 800°C under equilibrium, whereas little Ni dissolves in Si.

10.4.1.2 Crystal Structure and Volumetric Change

Silicide formation not only consumes Si but also induces volume change. The latter may give rise to mechanical stresses to be treated in Section 10.4.4. The amounts of Si consumption and volume change depend on the silicide phase formed. They can easily be calculated based on the crystal structure and unit cell dimensions of the silicides of interest which can be found in several sources such as the *Joint Committee on Powder Diffraction Standard Database* (JCPDS) published by the International Center for Diffraction Data [71] or *Pearson's Handbook of Crystallographic Data for Intermetallic Phase*s published by American Society for Metals [72]. For the Ni–Si system, the relevant information extracted from such sources is given in columns 2 through 7 of Table 10.2. In detail, the six silicide phases stable at room temperature (JCPDS card numbers: 06-690 Ni_3Si, 31-638 $Ni_{31}Si_{12}$, 03-943 Ni_2Si, 17-0881 Ni_3Si_2, 38-0844 NiSi, and 43-989 $NiSi_2$) are listed along with pure Ni (4-850) and pure Si (27-1402). The crystal structure and space group (column 2) as well as unit cell dimensions (column 3) lead to calculations of the unit volume (V, i.e., the volume of a unit cell) and the chemical formula unit per unit volume in columns 4 and 5, respectively. The following two columns present the average volume per Ni atom (V_{Ni}) and per Si atom (V_{Si}) for each of the crystal structures. They present, in fact the inverse of the atomic density for each element in the compound. From the data in column 6, the ratio of the silicide thickness grown ($t_{silicide}$) to the thickness of the Ni film deposited (t_{Ni}) is readily obtained; the volume per Ni atom for the silicides divided by the volume per Ni atom for pure Ni yields the ratio of $t_{silicide}$–t_{Ni} in column 8. Similarly, the data in column 7 lead to the ratio of $t_{silicide}$ to the thickness of the Si consumed (t_{Si}) in column 9. Finally, the amounts of Si consumed can be compared to the amounts of Ni deposited in column 10. The numbers given here usually represent data for relatively thick films or powders and are therefore more representative of bulk samples. In thin films, some variations can be expected depending on grain boundaries, defects, impurities and strains. The grain size in thin films, can change during formation, the impurities can segregate, some defects can either be generated or annihilated by the phase formation, and the strain can vary from film to film or grain to grain. Thus, the ratio of the silicide film thickness to the Ni thickness could deviate from the tabulated values. However, from our experience, this ratio is usually close to the expected ratio from bulk samples.

10.4.1.3 Properties of Ni Silicides

In Table 10.3, some physical properties of the six Ni silicide phases stable at room temperature are presented: they are, specific resistivity, melting point or transformation temperature, average Young's modulus, average coefficient of thermal expansion (CTE) and enthalpy of formation. NiSi is the desired phase for contacts in devices because of its low resistivity. The melting point of the silicides provides an indication of the morphological stability of thin films. A low melting point most likely leads to substantial atomic diffusion of the constituents already at a low temperature. As a result, degradation of the silicide film occurs at low temperatures. Note that the metal-rich phases show considerably higher melting points than NiSi and $NiSi_2$ except Ni_3Si_2. This latter phase shows low transformation temperatures with first a polymorphic transformation at about 830°C and a peritectoid transformation to NiSi and the high temperature θ phase. Since these temperatures do not refer to the presence of liquid, the morphological stability of Ni_3Si_2 may not be much poorer than other phases.

The elastic constants and the CTE shown are thin film averages. Combined, they allow for the determination of average strain and stresses in thin Ni-silicide films. While it is considered that NiSi should be advantageous over $CoSi_2$ or $TiSi_2$ in terms of stress reduction in devices [73,74], it will be shown in Section 10.4.5 that the large anisotropy with the CTE of NiSi may lead to drastic variations of the stress in individual NiSi grains from tensile to compressive as the grain orientation changes. These large variations in local stresses could be detrimental to device performance and reliability. In general,

TABLE 10.2 Crystallographic Properties of Ni, Ni–Si Phases and Si

Phase	Crystal Structure/Space Group	Lattice Constants (Å)	Unit Volume V (Å^3)	Formula Unit Per V	V_{Ni} (Å^3)	V_{Si} (Å^3)	$t_{silicide}/t_{Ni}$	$t_{silicide}/t_{Si}$	t_{Si}/t_{Ni}
Ni	Cubic/Fm3m	$a=3.523$	43.76	4	10.94	—	—	—	—
Ni_3Si	Cubic/Pm3m	$a=3.505$	43.08	1	14.36	43.08	1.31	2.15	0.61
$Ni_{31}Si_{12}$	Hexagonal/P321	$a=6.671$, $c=12.288$	1421	3	15.28	39.46	1.40	1.97	0.71
Ni_2Si	Orthorhombic/Pcmn	$a=7.39$, $b=9.90$; $c=7.03$	514.3	16	16.07	32.15	1.47	1.61	0.91
Ni_3Si_2	Orthorhombic/Cmc21	$a=12.22$; $b=10.80$; $c=6.924$	919.5	16	19.16	28.73	1.75	1.44	1.22
NiSi	Orthorhombic/Pnma	$a=5.233$; $b=3.258$; $c=5.659$	96.48	4	24.12	24.12	2.20	1.21	1.83
$NiSi_2$	Cubic/Fm3m	$a=5.416$	158.0	4	39.50	19.75	3.61	0.987	3.66
Si	Cubic/Fd-3m	$a=5.430$	160.1	8	—	20.01	—	—	—

Source: Adapted from Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004. With permission.

TABLE 10.3 Physical Properties of Ni, Ni–Si Phases and Si

Phase	Resistivity (μΩ cm)	Transformation Temperature/ Melting Point (°C)	Avg Young's Modulus (GPa)	Avg CTE for Thin Film	Enthalpy of Formation (kJ/mol)
Ni	7–10	–/1455	200	13.4	—
Ni_3Si	80–90	1035/1170	139	9.0	149
$Ni_{31}Si_{12}$	90–150	–/1242	177	—	1850
Ni_2Si	24–30	1255/1306	161	16.5	132–143
Ni_3Si_2	60–70	830/845	167	—	224–232
NiSi	10.5–18	–/992	132	∼12	85–90
$NiSi_2$	34–50	981/993		3.9–5.4	87–94
Si	Dopant dependent	–/1414	130–187	2.60	—

Note that the coefficients of thermal expansion (CTE) and the Young's modulus represent an average in polycrystalline thin films. These quantities may vary significantly with crystalline direction.

Source: Adapted from Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004. With permission.

the elastic characteristics of thin films are expected to vary with film texture since the properties of a solid crystalline substance may vary significantly along different crystallographic directions. The Young's modulus for any direction in a crystal as well as the shear stress can be extracted from the second order stiffness tensor that relates strain to stress [59,75,76]. In a cubic material such as Si, the Young's modulus varies from 130 to 187 GPa depending on the crystal orientation. It can therefore be expected that the values given in Table 10.3 are very sensitive to film texture and can vary from one reference to the next.

The enthalpy of formation is useful in evaluating the magnitude of the driving force for a reaction. As will be discussed later, from the difference in enthalpies one can evaluate if nucleation may be a controlling mechanism in the formation of a specific compound. The values of enthalpy of formation for the Ni-silicides are listed in kilojoule per mole of compound. Defined this way, the very high value for $Ni_{31}Si_{12}$ simply reflects the presence of a large number of atoms in one formula unit (43).

10.4.1.4 Low Silicon Consumption during NiSi Formation

Since $CoSi_2$ and $NiSi_2$ share the same crystallographic structure and have almost identical unit cell dimensions [71,72], it is valid to discuss $CoSi_2$ by referring to the values for $NiSi_2$ in Table 10.2. Comparing with $CoSi_2$, two factors contribute to a relatively low Si consumption during NiSi formation: (1) a lower resistivity of NiSi (15 μΩ cm vs. 18 μΩ cm for $CoSi_2$) and (2) a lower density of Si in NiSi (column 7 in Table 10.2). To obtain a given sheet resistance, the lower resistivity with NiSi indicates a thinner NiSi film needed as compared to a $CoSi_2$ film. Less Si is consumed as a result. The lower density of Si in NiSi further reduces the Si consumption to reach the specified sheet resistance, which is schematically shown in Figure 10.13. A reduction of Si consumption by more than 30% is calculated. This of course is dependent on the selected resistivities for the thin film. The gain in conductivity that is valid for bulk samples may not be as large in thin films. Moreover, since the NiSi films are typically thinner than their $CoSi_2$ counterparts, surface scattering effects could significantly affect the sheet resistance of such films.

10.4.2 Reactive Phase Formation

10.4.2.1 Thermal Budget

The formation of the desired low-resistivity NiSi occurs at a considerably lower temperature than the preceding widely used C54-$TiSi_2$ and $CoSi_2$ [22,74,77,78]. To demonstrate this low temperature formation, an example of in situ resistance and diffraction measurements on a 15 nm thick Ni film

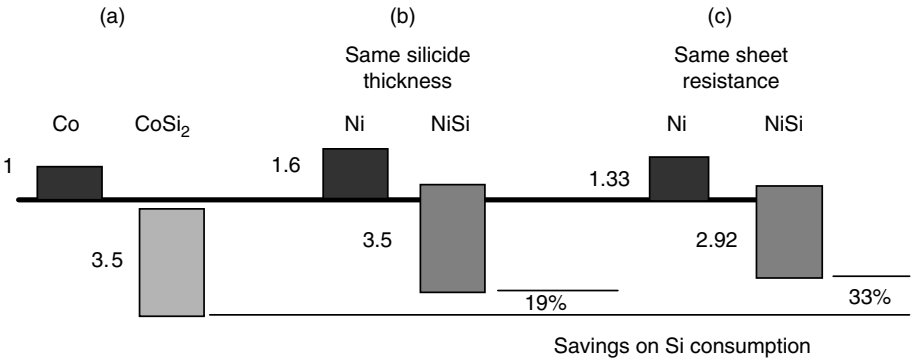


FIGURE 10.13 Schematic illustration of the reduction in Si consumption when using NiSi as a replacement for $CoSi_2$. (Reproduced from Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004. With permission.)

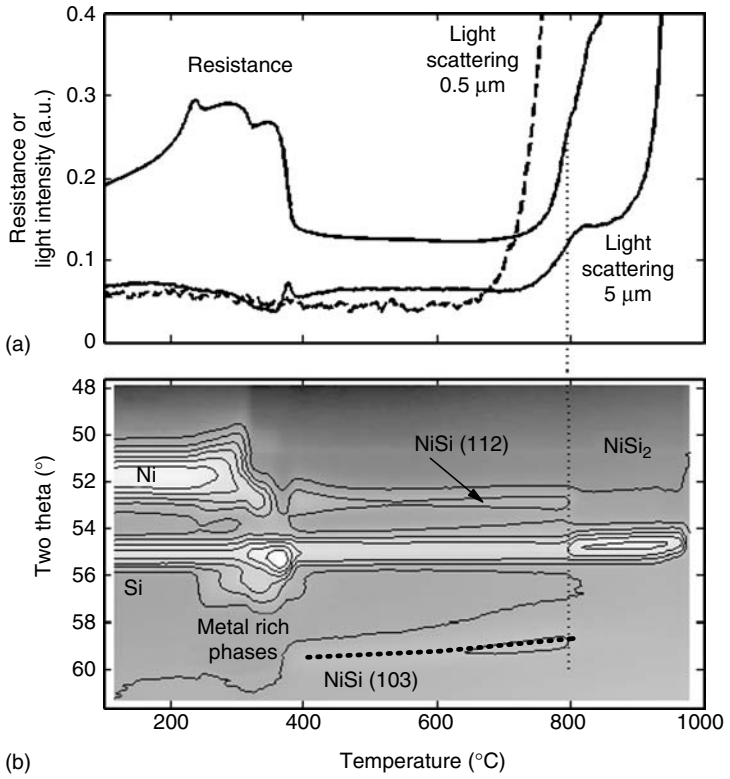


FIGURE 10.14 (a) Resistance and light scattering from 0.5 to 5 μm length scales together with (b) x-ray diffraction measurements performed in situ during annealing (3°C/s) of a 15-nm thick Ni film deposited on *p*-type poly-Si. (Reproduced from Lavoie, C., d'Heurle, F. M., Detavernier, C., and Cabral, C., *Microelectron. Eng.* 70, 144, 2003; Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004. With permission.)

deposited on undoped poly-Si is shown in Figure 10.14. In the upper panel of Figure 10.14, the resistance trace was measured while the substrate temperature was raised at a constant rate of 3°C/s in nitrogen. The film resistance is continuously modified as the temperature increases. Since the silicide phases have distinct resistivity values as seen in Table 10.3, the resistance variation can be a signature of the successive formation of the different silicide phases. Phase identification using XRD in the lower panel of Figure 10.14, confirms that the low resistance plateau spanning from 350 to 650°C corresponds to the formation of NiSi. A detailed analysis of the diffraction data will be given below using a three-dimensional image. A sharp increase in resistance occurring above 700°C can be caused by either agglomeration or $NiSi_2$ formation, two different degradation mechanisms to be discussed later. This early thermal degradation of the sheet resistance emphasizes the importance of reducing the subsequent anneals after the NiSi formation.

10.4.2.2 Complex Phase Sequence

The complex phase diagram for the Ni–Si binary system in Figure 10.12 indeed leads to a complex phase formation behavior recently discovered by a combined use of elaborated analysis techniques such as XRD using intensive synchrotron radiation, film stress and surface light scattering all performed in situ when the substrate temperature is increased [47,79–82]. Traditional studies of the interaction of Ni films with Si substrates, using relatively thick films ($\sim$100 nm Ni) and isothermal anneals followed by ex situ characterization, revealed only the sequential formation of Ni_2Si, NiSi, and $NiSi_2$ [59,83–99]. For samples prepared by standard device manufacturing processes and for film thicknesses relevant to the state-of-the-art CMOS technology, several phases, i.e., $Ni_{31}Si_{12}$, Ni_2Si, and Ni_3Si_2, appear in sequence or parallel during RTA prior to the NiSi formation [47,79–82]. Recent publications also report local epitaxy of discrete $NiSi_2$ grains, in the form of $NiSi_2$ inverted pyramids, early in the phase sequence which are subsequently consumed by the formation of Ni_2Si or NiSi [88,100]. In small openings below 0.6 μm in diameter or side length, epitaxial $NiSi_2$ was found already at 400°C [101]. It may be recalled that besides sharing similar crystallographic structures, the unit cell dimension of $NiSi_2$ at room temperature (5.416 Å) is very close to that of Si (5.430 Å). This small difference will be reduced as the temperature is increased since the thermal expansion of $NiSi_2$ is larger than that of Si, see Table 10.3 [59,102]. The formation of $NiSi_2$ inverted pyramids seems to be very sensitive to the amounts of oxygen at the interface and a subtle difference in interfacial oxygen may have hampered the attempt to reproduce the observations [22].

The remarkable advantage of in situ characterization has already been emphasized with multiple examples through the current chapter using two-dimensional rendering for the diffraction part of the experiment. A three-dimensional presentation can be more revealing of the phase formations. In Figure 10.15, the data from the same anneal as in Figure 10.14 (15 nm Ni/*p*-type doped poly-Si) is presented in such form [19,20,47]. The two-dimensional presentation in Figure 10.14b represents a top view of the three-dimensional presentation in Figure 10.15.

Two peaks are present at low temperatures in the selected 2θ window (50°–60°). The first one near 52°C is the (111) diffraction peak of Ni. The second, slightly above 55°C, is the (220) diffraction line from the poly-Si substrate. As the temperature increases to 300°C, the Ni peak starts to decrease in intensity while another peak of high intensity appears at about the same position as the Si (220). Considering the metal-rich phases alone, this peak position matches the (350) spacing of Ni_3Si_2 [55]. The onset temperature of this phase is better revealed in the two-dimensional figure to be around 370°C. Additional details regarding the identification of the metal-rich phases were published earlier and supported with limited TEM and other XRD data [47,80,103]. While the Ni_3Si_2 matches the XRD peaks observed, one can not exclude possibilities that these peaks be related to unknown metastable phases or to a few strained NiSi grains that would be textured differently than the NiSi forming at higher temperatures. From then on, in the chapter, this phase will be preferentially referred to as Ni_3Si_2. When the temperature reaches 400°C, the peak from the metal-rich phase has disappeared and two low intensity peaks from the NiSi phase can be observed: (103) just below 60° and (112) at about 53°. Note that there is a clear decrease of the Si (220) peak intensity, which marks the consumption of the poly-Si layer during Ni silicide formation. At about

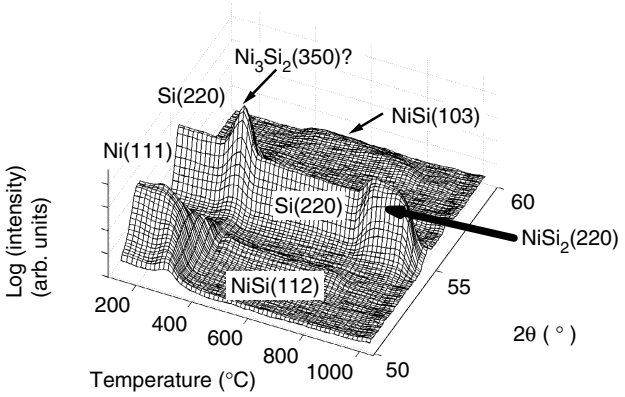


FIGURE 10.15 3D rendering of the in situ x-ray diffraction measurements in the lower panel of Figure 10.14.

800°C, the two weak NiSi peaks disappear as the peak intensity of $NiSi_2$ (220) increases. This phase forms relatively quickly and is stable over a temperature range of less than 200°C as both the poly-Si and $NiSi_2$ intensities decrease drastically upon reaching the eutectic melting temperature of 966°C. The decrease in $NiSi_2$ intensity is a direct consequence of melting. However, the disappearance of the poly-Si peak is more surprising and could be the result of texturing in the poly-Si layer in the presence of the eutectic liquid.

What has not yet been detailed is the in situ light scattering for surface roughness analysis [104] performed simultaneously with the diffraction and resistance measurements. The results are shown in the upper panel of Figure 10.14 along with the resistance curve. The light scattering signal from the 0.5 μm scale remains low throughout the formation of all metal-rich and the NiSi phases, up to a temperature above 600°C. In contrast, a significant change in the 5-μm signal occurs slightly below 400°C. A local maximum at this length scale is observed at a temperature corresponding precisely to the sharp decrease in resistance and disappearance of the intense XRD peak that could be the Ni_3Si_2 (350). The large increase in light scattering signal at high temperatures does not begin until slightly below 700°C for the 0.5-μm signal and about 750°C for the 5 μm one. The 0.5 μm signal may provide some unique feature of the NiSi surface, while the 5 μm signal gives a temperature coinciding well with the onset temperature for a large increase in sheet resistance. The formation of the high-resistivity $NiSi_2$ does not take place until 800°C according to the XRD data. The light scattering signals registered are therefore indicative of surface roughening starting before the transformation from NiSi to $NiSi_2$.

Phase identification relying on XRD requires caution. The JCPDS files present data for unstrained powders typically at room temperature. Thermal expansion (Table 10.3), stress build-up and strain relaxation, impurity incorporation and precipitation during phase formation can all affect and complicate the identification of phases [57,81,82,105–108]. It is hence of interest and importance to combine different analysis techniques to yield complementary information. It is nevertheless clear that multiple phases form prior to the appearance of the low-resistivity NiSi phase. As the temperature windows over which each of these phases exists overlap significantly, multiple phases (>2) coexist in the growing film.

Since the phase sequence may influence the development of mechanical stresses in the thin film structure [82], a few words are given to the first metal-rich phase formed at low temperatures before we turn our attention to the next topic. During the phase formation of Ni silicides, the Ni_3Si phase may be the first silicide to form. This phase is however not easily distinguishable using the current XRD settings. If it did form, only a very weak shoulder peak on the Ni (111) peak would appear since Ni_3Si and Ni are both cubic with similar lattice constants. The first phase formed that can be clearly identified by XRD is a strained Ni_2Si phase. The formation of this first phase was carefully analyzed with isothermal anneals at

low temperature to help understand the formation mechanisms and formation kinetics [22,82]. From these studies, it is clear that the early formation of this strained Ni_2Si phase originates from the crystallization of an amorphous layer and that the strain within this crystallized Ni_2Si layer only relaxes when the Ni layer is completely consumed. More importantly, this phase coexists with the tentative Ni_3Si_2, which makes a transient appearance before the full consumption of the Ni layer [109]. Analysis of the first phase to form at low temperatures also has bearing on practical applications, since most tooling systems used in the microelectronics industry go through preheat cycles that are of low temperature nature.

10.4.2.3 Effect of Dopants on Phase Formation

In situ XRD is perhaps the most powerful means for demonstration of the complexity of the low temperature phase formation and its dependence on dopant type: Figure 10.16 shows two top-view diffraction patterns obtained during annealing 15 nm thick Ni films deposited on n- and *p*-type doped SOI substrates with a thin surface Si layer of 100 nm thickness on a 1200 nm thick buried oxide. Such substrates are currently being used in the semiconductor industry for multiple purposes including ease of device making and suppression of parasitic components in a circuit. For the *n*-type substrate, phosphorous was implanted to a dose of 5×10^{15} cm^{-2} at an energy of 12 keV while for the *p*-type substrates, boron was implanted to a dose of 3.5×10^{15} cm^{-2} at an energy of 8 keV. The anneals were performed at 3°C/s in purified He. The sequence of phase formation described earlier for a poly-Si substrate is clearly seen in the figure without the interference of the poly-Si (220) peak. An important observation for actual device applications is the phase formation occurring at temperatures 50°C higher

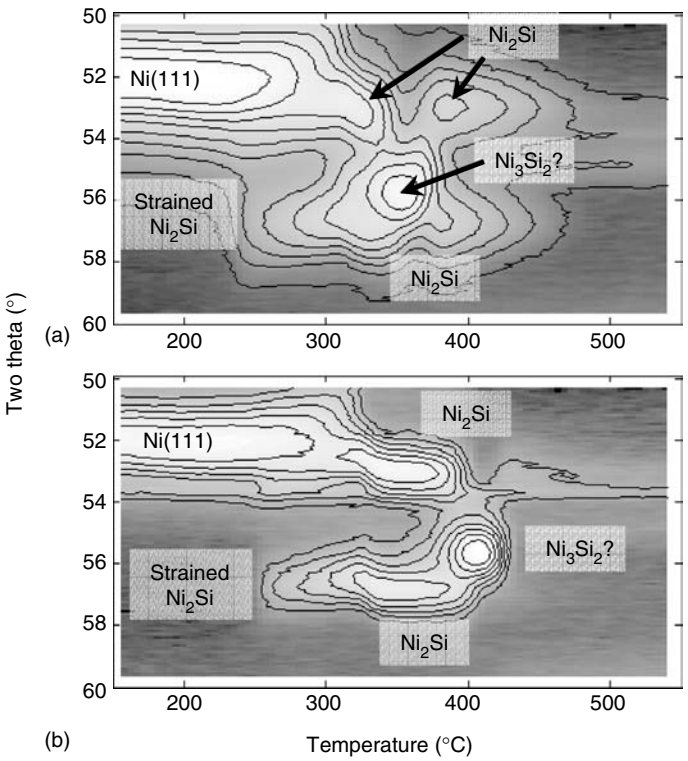


FIGURE 10.16 In situ x-ray diffraction measurements during annealing of a 15-nm Ni film deposited on heavily (a) *p*-type and (b) *n*-type doped SOI substrates. (Reproduced from Lavoie, C., d'Heurle, F. M., Detavernier, and C., Cabral, C., *Microelectron. Eng.* 70, 144, 2003 (corrected) and Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004. With permission.)

on the *n*-type substrate than on the *p*-type substrate. The formation temperatures on the *p*-type substrate are similar to those on undoped poly-Si in Figure 10.14. An annealing at too low a temperature may lead to different silicides formed and thus different resistance attained in the *p*- and *n*-type regions of a CMOS circuit.

It is unclear at this point how dopants influence silicide formation. Compound formation involving the dopants could be a cause, which received a great deal of attention 10–25 years ago when studying the Ti and Co silicide formation [110,111]. Dopant diffusion in the silicide films [112] is another factor influencing the phase formation. Consequently, Ti–Si-dopant and Co–Si-dopant ternary systems are very well understood and first estimations of the ternary phase diagrams of Ni–Si with various dopants are the main sources of information available [110]. The estimate suggests that the borides of Ni would interact with Si whereas the arsenides or phosphides of Ni could be stable in the presence of Si. While the ternary phase diagrams are all tentative, they seem to be supportive to the experimental observations of *n*-type dopants delaying the phase formation of Ni-silicides.

A high doping concentration in the Si at the interface between NiSi and Si is crucial for achieving a low contact resistance [5,110]. Several studies of dopant redistribution during silicide formation show that both phosphorous and arsenic tend to partially segregate at the interface between NiSi and Si [113–115]. Although only few studies compare the contact resistance of $TiSi_2$ or $CoSi_2$ and NiSi under similar conditions, most of them point towards a lower contact resistance for the NiSi contacts [5,74,110,116–119]. Furthermore, dopant diffusion in the Si can be affected by the silicide formation through injection of point defects, an effect confirmed for the formation of either $TiSi_2$ or $CoSi_2$ [120–122]. The diffusion of B and P is mainly mediated through interstitials and that of As and Sb through vacancies [123]. By monitoring the movement of these dopant elements, an increase in vacancy concentration and a decrease in interstitial concentration during the formation of $TiSi_2$ and $CoSi_2$ were concluded [120–122]. It is known [124,125] that the formation of all Ti-silicide phases as well as CoSi occurs via Si diffusion while that of Co_2Si and $CoSi_2$ through Co diffusion. Although this injection of point defects was shown to be independent of diffusion species for the Co and Ti disilicide formation [120], the fact that all Ni-silicide phases form through Ni diffusion could reduce the concentration of vacancies in the Si substrate. If so, the final junction profile under NiSi could be significantly different from that under $TiSi_2$ or $CoSi_2$. However, it is important to recall that the formation temperature used for these three phases differs substantially from 850 to 900°C for C54-$TiSi_2$, to 700°C–750°C for $CoSi_2$, and to 450°C–500°C for NiSi [22,74,77,111].

10.4.3 Formation Mechanism

10.4.3.1 Diffusion Controlled Formation

Among the most appreciated advantages with NiSi contacts are smooth surface and interface, in addition to the already mentioned low consumption of Si from the substrate. The smoothness is a direct result of the diffusion-controlled formation and growth of NiSi [57,89,92,107]; upon annealing, nucleation of NiSi at the interface between Si and the precursor silicide phase is easy, resulting in a high density of NiSi nuclei across the interface. This high density of nuclei leads to the formation of a continuous NiSi layer that thickens (i.e., grows) through transportation of Ni atoms across the growing NiSi [92]. As a result, the growth of NiSi is nearly planar with its reaction front moving uniformly across the interface following a parabolic relation where the thickness square is proportional to the time [12,125]. A diffusion-controlled reaction is in sharp contrast to a nucleation-controlled reaction as in the formation of C54-$TiSi_2$ [25,26], $CoSi_2$ [56] or $NiSi_2$ [126] where the reaction evolves non-uniformly and rapidly, typically above a critical temperature, and leads to some characteristic roughness [57].

Three factors determine the nucleation of a new phase [57]: the amplitude of the change in free energy per unit volume Δg for the formation of a nucleus, the amplitude of the change in surface energy $\Delta\sigma$ associated with the creation of the nucleus, and the atomic diffusion to make the nucleation

possible. The rate of nucleation R can be expressed as,

$$R \propto \exp\left(-\frac{\Delta G^*}{R_B T}\right) \times \exp\left(-\frac{Q}{R_B T}\right) \tag{10.1}$$

where, R_B is the Boltzmann constant and T temperature in Kelvin. The first exponential term on the right-hand side is related to thermodynamic quantities and represents the probability of nucleation or the density of nuclei of critical size (assuming equilibrium conditions). The second exponential term represents a kinetic term with Q as the activation energy for growth. ΔG^* represents the energy barrier height to overcome in order for nucleation to proceed. It is given by:

$$\Delta G^* = \frac{4}{27}\frac{b^3}{a^2}\frac{\Delta\sigma^3}{\Delta g^2} \tag{10.2}$$

with a and b as two geometrical constants determined by the detailed shape of the nucleus. From Equation 10.1 and Equation 10.2, for formations where Δg values (per unit volume) are small and consequently ΔG^* large, nucleation becomes difficult and rate limiting.

While multiple factors contribute to the free energy, to the first order, an idea of the type of reaction can be gained by following the variations in the enthalpy of formation. For solid-state reactions, enthalpy provides a good approximation of free energy as long as the difference in enthalpy of formation between the products and the reactants is large. Caution should be exercised when this difference is small as they can be different, one knows however that small enthalpy changes correspond also to "small" free energy changes. For Ni silicide phases, it has been reported that nucleation is only important for the formation of $NiSi_2$. These conclusions were reached when only three phases were observed, Ni_2Si, NiSi, and $NiSi_2$. They need to be revisited since as shown earlier the phase formation sequence is more complex in current thin films: Ni_2Si forms first, followed by the formation of a transient phase (Ni_3Si_2?) and coexistence with the Ni_2Si and Ni, after which, this extra phase is consumed by the formation of NiSi (and possibly Ni_2Si again [82,103]), further NiSi formation occurs at the expense of the remaining Ni_2Si and finally $NiSi_2$ forms from NiSi and Si. The appearance of the Ni-rich phases in the sequence may affect the stress development in the formed silicide thin films [82,127]. A careful exercise with the calculation of the energy difference at each step of the phase formation sequence indicates [22] that to the extent that the enthalpy or free energy values can be trusted, only the first phase richest in Ni, Ni_2Si in the previous phase sequence and in the current one, would form by a diffusion controlled process. The NiSi formation from Ni_2Si is also diffusion controlled with a relatively large change in enthalpy of formation. The inclusion of a new precursor phase Ni_3Si_2 (or any metastable phase for that matter) before NiSi in the current sequence has considerably reduced the change in enthalpy of formation. Potentially, the nucleation could become more influential in forming the NiSi. Since experimentally thin NiSi films are extremely smooth in comparison with the disilicides of Ti, Co, and Ni, the formation of NiSi in the current films is still mainly controlled by diffusion. If factors such as changes in the density of defects or variation in mixing entropy in the presence of alloying elements become pronounced, the free energy of formation could be decreased to the point where nucleation may become an issue and the film could become significantly rough.

10.4.3.2 Dominant Diffusing Species: Reduction of Bridging and Kirkendall Voiding

It is important to recall that Ni is the dominant diffusing species during the formation of the Ni silicides [125]. This fact leads to two obvious advantages with regard to device manufacturing with the self-aligned process. The first one is a reduced risk of electrical shorts between the gate and source-drain electrodes, i.e., bridging, even for aggressively scaled CMOS devices. This can be achieved if the temperature of formation is kept sufficiently low so that Si is not significantly mobile. With a low diffusivity for Si atoms, the formation of a silicide above sidewalls or oxide areas that would resist the selective etch is minimized.

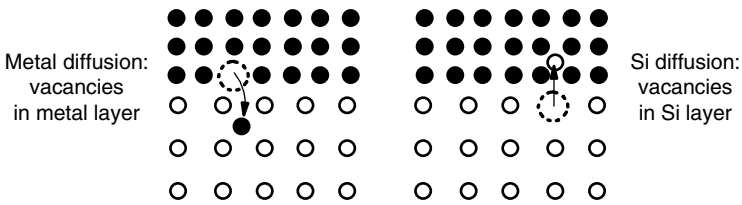


FIGURE 10.17 Schematic illustration of the importance of the diffusing species in determining where void formation is likely to occur when reacting a metal film with a thin Si layer (e.g., on SOI wafers). (Reproduced from Lavoie, C., d'Heurle, F. M., Detavernier, C., and Cabral, C., *Microelectron. Eng.* 70, 144, 2003; Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004. With permission.)

The second advantage of Ni diffusion is presented in Figure 10.17. The vacancies generated by the departure of Ni are now mainly located in the deposited Ni layer as compared to the vacancies formed in the Si when the Si is diffusing during the formation of CoSi [22]. The consequences of vacancies forming in the metal layer, although not obvious at first, are of critical importance in small dimensions with limited supply of materials. The diffusivity of vacancies is very high. If they are generated in bulk substrates because of Si being the dominant diffusing species, as is the case for the formation of $TiSi_2$ or CoSi, they can rapidly equilibrate and be distributed throughout the wafer thickness. However, in many cases of bulk diffusion couples the vacancies can result in a development of diffusional porosity also called Kirkendall voiding [75,128–130]. In the case of silicide formation on SOI wafers and poly-Si gates, the Si volume that is to accommodate the vacancies is not so much larger than the volume of the vacancies generated during formation. Under such conditions, voids may be much easier to form. The presence of grain boundaries in poly-Si gates is anticipated to worsen the void formation since they act as sinks for the vacancies. When metal diffusion is predominant and the vacancies are generated in the metal layer, the morphology of this layer after reaction is of little interest since it is removed by selective etching. This is believed to be an important reason why $CoSi_2$ shows limitations in small poly-Si dimensions whereas NiSi does not [22,68, 131–135] (See Figure 10.11).

10.4.3.3 Disadvantages of Rapid Diffusion at Low Temperature

The rapid diffusion of Ni has led to the so-called "reversed fine line effect" where lower resistance is found in narrower dimensions [65,67,68,74,136]. When the annealing temperature is such that Ni is allowed to diffuse over a distance longer than the thickness of the film, more NiSi grows towards the edge(s) than in the middle of a narrow line. When the line is sufficiently narrow, the silicide thickness in the entire line is substantially thicker than in wider lines. Consequently, the resistance in narrower lines is lower than in wider lines [22,74]. The extra Ni atoms above the spacer or on the shallow trench isolation (STI) that are within the diffusion distance can increase the amount of reaction close to the Si line edges. It is apparent that this small percentage for large dimension structures can become significant in narrow lines. A liable solution to the fine line effect is to use a two-step annealing process [65]. The first annealing is controlled at low temperatures around 300°C to limit the lateral diffusion of Ni while inducing the self-aligned silicide formation. The second annealing at about 500°C, after a selective removal of the unreacted Ni, ensures the formation of the low-resistivity NiSi. A relatively high temperature anneal may also be needed to ensure that all Ni atoms that have diffused into Si have become a part of the silicide contact. Interstitial Ni atoms diffuse extremely rapidly in Si and can create deep level defects detrimental to device performance. Literature reports suggest that annealing at a temperature as high as 500°C may be necessary to eliminate carrier trapping by Ni atoms as measured by deep-level transient spectroscopy (DLTS) [65,137,138]. It is unclear however, if the high temperature simply drive the Ni atoms away from the probing area of DLTS.

Once a quality NiSi film is formed, the next challenge is to retain its integrity as a continuous film during further temperature exposure. Two important and unexpected characteristics of the NiSi as a crystalline material itself or as a thin film grown on Si substrate can affect the integrity of NiSi thin films at elevated temperatures. The first one is the very large anisotropy in CTE [139–141] and the second one is a new kind of texture observed in NiSi films on single crystal Si (sc-Si) substrates [142]. Treatment of these characteristics is delayed to Section 10.4.5 and Section 10.4.6.

10.4.4 Challenges and Progress

It is crucial for a NiSi thin film to retain its integrity and low resistance even upon further anneals after the silicide formation. There are two causes responsible for a NiSi film to degrade at high temperatures. The first cause is the formation of the high-resistivity phase $NiSi_2$ at the expense of NiSi, which also yields a considerably roughened interface [12,57,126]. The second cause is morphological degradation of the film through grain boundary grooving and agglomeration. In the following, these are discussed followed by an account of progress made to improve the integrity.

10.4.4.1 Formation of $NiSi_2$

In the presence of excess Si, a formed NiSi should transform to $NiSi_2$ according to the Ni–Si phase diagram in Figure 10.12. The transformation usually does not occur until about 800°C (see Figure 10.14) because of a very small change in free energy associated with this phase formation therefore a high nucleation barrier for the formation of $NiSi_2$ [57]. The sudden appearance of the $NiSi_2$ phase at high temperature is a typical signature of a nucleation-controlled formation. In addition to a high resistivity and a roughened interface, the main disadvantages with the appearance of this phase are (1) twice as much consumption of Si in comparison with the formation of NiSi, and (2) the tendency of epitaxy [143–151]. A more Si consumption is especially hazardous for contacts formed on SOI substrates. Note here that although the $NiSi_2$ formation leads to a two-fold increase in thickness (from NiSi thickness), the resistance of the film itself will still increase by 50% since the increase in resistivity is three times.

The crystal structures (face-centered cubic) and dimensions of the unit cells for $NiSi_2$ (5.416 Å) and Si (5.430 Å) are very similar. The difference in lattice constant becomes smaller with increasing temperature since the thermal expansion of Si is smaller than that of $NiSi_2$. When epitaxy of $NiSi_2$ occurs, the growth of these grains will not be limited to the open contact areas but will tend to simply follow the crystal directions of the underlying substrate leading, for example, to encroachment under sidewall spacers or to the formation of epitaxial facets that increase interface roughness. This was initially observed for epitaxial grains of $CoSi_2$ [48,58,152,153] and was one of the reasons why the industry shied away from the otherwise promising epitaxial silicides [144,154]. Epitaxy of $NiSi_2$ on Si tends to be easier than that of $CoSi_2$, the exact cause for the different behaviors is unknown although the lattice mismatch between $NiSi_2$ and Si is smaller (0.4%) than that of $CoSi_2$ and Si (1.3%).

10.4.4.2 Morphological Stability

The low melting point of NiSi indicates a high diffusivity of constituent atoms, Si and Ni, which is detrimental for the morphological stability of NiSi films. The mechanisms behind the morphological instability of NiSi on Si are different depending on the substrate type [23,155–157]: grain growth in poly-Si for NiSi on poly-Si *vs.* grain grooving of NiSi on sc-Si. Both processes, grain growth and grain grooving, are driven by a minimization of free energy of the system in question. The temperature for grain growth in poly-Si (i.e., degradation temperature) was found [158,159] to scale with the melting temperature of the silicide layer atop, with an approximate relation as $T^{\text{poly-Si}}_{\text{grain-growth}} = 0.6 \times T^{\text{silicide}}_{\text{melting}}$. Although the degradation temperature might also be represented as to scale with the formation temperature of silicides, a better correlation was found [160] on the basis of a careful study on Ni-, Pd-, Pt-, Co-, Ti- and Cr-silicides: it is related to the deformation temperature of silicides by $T^{\text{poly-Si}}_{\text{grain-growth}} = T^{\text{silicide}}_{\text{deformation}} + 100$.

The high diffusivity of Si and Ni in NiSi as a consequence of the low melting point of NiSi is kinetically favorable for grain grooving. The minimization of total energy at high temperature leads to separation of the film into more spherical islands for which there is an overall decrease in surface or interface areas for the same volume of NiSi and therefore a decrease in surface and interface energies. In comparison with the formation of $NiSi_2$, the morphological instability becomes more predominant for thinner NiSi films that are more relevant to current and future CMOS technologies [22,88,157,161,162]. The evidence is the morphological degradation occurring without the formation of $NiSi_2$; as an example [147], for a 45 nm thick NiSi morphological degradation was observed to begin at 700°C while the formation of $NiSi_2$ did not occur until 800°C. For thicker films, the formation of $NiSi_2$ takes place at a temperature lower than that of the morphological [119,163–166].

It is generally true that morphological degradation of thin films occurs at lower temperature on poly-Si than on sc-Si. Large grained poly-Si layers would act more like sc-Si in this respect. The preparation of the initial poly-Si substrate, doped or undoped, high temperature processed or as deposited at relatively low temperature grain structures, can therefore influence the grain size in the poly-Si and thus the degradation mechanism and temperature of the thin layer. It has however been found that the agglomeration of NiSi on SOI substrates is peculiar. It would be expected that NiSi films on SOI substrates be more stable against agglomeration since the substrate is single crystal (equivalent to one extremely large grain). But the experimental observations showed [22,157] that while the degradation behaves similarly as on bulk sc-Si for NiSi films thicker than 35 nm, a different behavior is observed for very thin films. For the thinner films, the agglomeration rate on SOI is actually faster than on poly-Si substrates.

10.4.4.3 Improving NiSi Films

In recent years, several approaches have been found to reduce the tendency of NiSi films towards the formation of $NiSi_2$ as well as agglomeration. Among the methods specified, the addition of either Pt or Pd has been shown to push the formation of $NiSi_2$ to increased temperatures [22,167–175]. Other techniques have been suggested, such as implantation of nitrogen [126,176–179], hydrogen [164] or BF_2 [180,181] and the use of capping layers [182], which retard the agglomeration of films. It was shown that fluorine introduced through the implantation of BF_2 segregates to the NiSi/Si interface and retards agglomeration significantly [183–185a]. Recently, we also published an overview study of the effect of alloying more than 20 elements with Ni on the morphological stability of NiSi films [185b].

The addition of Pt or Pd in Ni either as an alloy or as an interposed layer works on the principles of fundamental thermodynamics; an analysis along this line resulting in a simplified ternary phase diagram for the Ni–Pt–Si system is found in [78]. In short, the phenomenon can be attributed to an extra free energy term due to the entropy of mixing in the ternary monosilicide solutions since PtSi as well as PdSi, are miscible with NiSi while ternary solid solutions of these disilicide do no exist since they are not found under equilibrium conditions. As discussed earlier, the free energy change Δg in Equation 10.2 for the transition from NiSi to $NiSi_2$ is already very small making it difficult. Any extra energy term in the monosilicide that further decreases Δg for the transition from the ternary monosilicides to $NiSi_2$ makes the formation of the later more difficult [62,69,167] In addition to raising the formation temperature of $NiSi_2$, Pt raises and widens the temperature window over which the metal-rich silicide phases are present [22,131]. Pt additions have been shown to improve of the morphological stability of the monosilicide film by as much as 150°C [131,170]. The improvement in morphological stability can be understood partly at least because of a higher melting point and a higher deformation temperature with PtSi than with NiSi [160]. Hence, it can be expected that the ternary monosilicide (NiPt)Si is more morphologically stable than NiSi. However, recent results imply that the presence of Pt may be detrimental to device performance as measured using DLTS [186].

Depending on the dose used, BF_2 implantation has been shown to increase the agglomeration temperature by up to 200°C [22,183]. This enhanced stability is attributed to the segregation of the fluorine at the NiSi/Si interface [183]. This 200°C rise in process window is extremely large and shows that either diffusion at the NiSi/Si interface or interfacial energy itself can be controlled. Apparently by

modifying the silicide through addition of elements or impurities, the properties can be adjusted to withstand the temperature exposure necessary for the processing of contacts and interconnections.

10.4.5 Anomalous Thermal Expansion of NiSi and Related Stress Effects

A close scrutiny of the diffraction peaks of NiSi in Figure 10.14 reveals a striking difference in the ways the (112) and (103) peaks shift as a function of increasing temperature. This behavior indicates an anisotropy of the thermal expansion of NiSi along its crystallographic axes. The anisotropy is known for bulk NiSi [139,140], and Figure 10.14 confirms its existence in NiSi thin films as well. The value for CTE determined from bending measurements [53] given in Table 10.3 represents an average comprising contributions from all possible combinations of crystallographic orientations. In aggressively scaled devices, one needs to understand what implications such an anisotropy would have, to the occurrence of local mechanical stresses in thin NiSi layers.

10.4.5.1 Unit Cell Dimensions and CTE for Bulk Samples

NiSi in an orthorhombic MnP structure has the room-temperature unit cell dimensions of $a = 0.5177$ nm, $b = 0.3325$ nm, and $c = 0.5616$ nm determined using XRD on polycrystalline bulk NiSi samples [140]. These values deviate slightly from a similar study but on NiSi single crystals [139]: $a = 0.51752$ nm, $b = 0.3321$ nm, $c = 0.56094$ nm. The small difference between the two sets of dimensions is within the experimental errors. Through measurements of the inter-planar lattice spacings of various crystallographic planes by means of XRD, the CTE as a function of temperature for the bulk samples were found to be strongly non-linear [140]. A third order polynomial was required to capture the temperature dependence of the unit cell dimensions. What is surprising was the observation of an unusual negative CTE with the shortest axis of the NiSi unit cell: it contracts with increasing temperature. A positive temperature dependence of CTE implies an increase in disorder and consequently an increase in volume. However, there is no fundamental reason, of thermodynamic origin or otherwise, that excludes a decrease in lattice parameter with increasing temperature. "Thermal contraction" is nevertheless extremely rare and several examples were listed in [22]. It should be noted that although the b axis of the NiSi unit cell contracts upon heating, the total volume of the unit cell actually increases with increasing temperature.

10.4.5.2 Unit Cell Dimensions and CTE for Thin Films

The unit cell dimensions for NiSi thin films deviate significantly from those for the bulk samples [85]: $a = 0.5233$ nm, $b = 0.3258$ nm, and $c = 0.5659$ nm, suggesting that NiSi thin films are strained. By tracking the variation of lattice spacings in situ while continuously annealing 20-nm thick NiSi films formed on poly-Si substrates, a large anisotropy of the CTE was also evident in thin films [141]. In particular, a very good match was found between bulk data and thin film behavior above 500°C implying a proper relaxation of the mechanical stresses above this temperature [81,82]. At lower temperatures, the variations in the unit cell dimensions with temperature were influenced by the unrelaxed strain.

The reason why the room-temperature values of a and c in thin films on Si substrates are significantly larger than the bulk values, whereas the b is slightly shorter in thin films than in bulk samples, is due to the thermal expansion mismatch between NiSi and Si as well as the anisotropic CTE of NiSi. Above 500°C, the mechanical stresses built in the thin films because of the volume change upon silicide formation (cf. Table 10.2) were properly relaxed [81,82]. Upon cooling back to room temperature, thermal stresses are built up as a result of the difference in CTE between film and substrate. A truly quantitative interpretation would require a precise knowledge of the strain-stress tensor for this material with undoubtedly peculiar elastic properties. A complication of the strong anisotropic thermal expansion of NiSi is the dependence of the room-temperature stress-strain state in the NiSi film on the grain orientation, the plane parallel to the NiSi/Si interface, the expansion coefficients in this plane and the stress-strain tensor. This can be further complicated by second-order effects such as those resulting from inter-grain stresses generated by the mechanical interaction between neighboring grains [187].

A qualitative account of the results of the room-temperature thin film values of *a*, *b*, and *c* can be found in [22]. The major points are that the *a* and *c* axes have large, positive expansion coefficients, averaging to about 40 ppm/°C, considerably greater than the value of 2.6 ppm/°C for Si. For the *b* axis, the coefficient is also large but *negative*, a contraction therefore occurs and follows a sigmoidal curve as the temperature increases, a behavior quite far from linear. According to the thermal expansion coefficients for NiSi thin films [141], an ideal unit cell that is embedded in a Si substrate and relaxed at 500°C, would be in tension in one direction (*c* and *a*) and in compression in the other (*b*) after cooling to room temperature. The results indicate that in the unknown stress-strain tensor the effect of compression along the *b* axis is important.

A consequence of these considerations is that the unit cell dimensions in the JCPDS record (38-0844) for NiSi, however faithfully they match room temperature measurements on thin films, are not equilibrium values. They correspond to films on Si substrates that suffer both from considerable local tensile or compressive stresses depending on grain orientation, and from not being equilibrium powder diffraction samples. Valid unit cell dimensions are found in bulk studies [139,140].

10.4.5.3 Consequences of CTE Anisotropy for Stress in Thin Films of NiSi

If one approximates the thermal expansion of NiSi as being linear as a function of temperature, one obtains values for the CTE of 42, -43 and 34 ppm/°C along the *a*, *b*, and *c* axis, respectively. When comparing the CTE values for the axes of the NiSi unit cell to known values for other silicides (typically 10 ppm/°C) and to the CTE for the Si substrate (2.6 ppm/°C), it is clear that the thermal expansion in NiSi is significantly larger. This seems in apparent contradiction with the relatively low stress that has been reported in NiSi films using wafer curvature methods [56,73]. The low average stress (as measured by wafer curvature) at room temperature is first of all related to the low stress relaxation temperature for NiSi. For most silicides, this stress relaxation temperature is similar to the typical growth temperature at which the phase forms in a reasonably short time in thin film experiments. Since the stress relaxation temperature for NiSi is much lower than that of $CoSi_2$ or $TiSi_2$, less thermal stress can build up while cooling the sample back to room temperature. Secondly, the rate at which, thermal stress builds up during cooling of the sample is determined by the product of the elastic modulus and the thermal expansion mismatch. The average bulk modulus for NiSi has been reported as 132 GPa in the literature [188]. However, this value should be used with caution, since the elastic properties of NiSi are undoubtedly also strongly anisotropic. In most cases, the bulk modulus is approximately inversely proportional to the thermal expansion coefficient (when positive), given that both quantities are related to the attractive part of the atomic potential within the solid: when the bulk modulus is large for a certain material, it is more difficult to increase the separation between atoms, and hence one expects a small tendency for the bond length to increase upon heating. This may explain why the slope of the increase in thermal stress when the NiSi layer is cooled down after reaction is similar to the slopes observed for other silicides, in spite of the significantly larger mismatch in thermal expansion for NiSi.

Although the average stress in NiSi films is low, the strong anisotropy of the CTE will cause significant localized stresses between neighboring grains. Indeed, XRD observations indicate that while some grains are under tension, others grains within the same film are experiencing compressive stress [22]. The strain energy density within a grain is determined by the product of stress and strain. As a consequence, the strain energy density within each individual grain of an anisotropic material is dependent on the orientation of the grain with respect to the substrate. Anisotropy in the modulus for metal films (e.g., Cu [189,190]) has been proposed as a possible texture selection mechanism since differences in strain energy between neighboring grains may provide a driving force for favorably oriented grains to grow faster, thereby influencing the texture of the film after annealing. However, the differences in strain energy between neighboring grains do not seem to provide a sufficient driving force to significantly influence the measured texture of the NiSi film since experiments on poly-Si or amorphous Si did not provide evidence of strong preferential orientation.

10.4.6 Texture Development in NiSi Films

Performing XRD θ–2θ scans is a common procedure in studying phase formation. When the NiSi formation on Si(001) is analyzed, all major diffraction peaks that are listed in the JCPDS file obtained from a powder sample are usually detected. An example of such a diffractogram is shown in Figure 10.18 where similar NiSi formation on a poly-Si substrate is also included for comparison. This observation unfortunately often gives a false impression that the NiSi film is randomly oriented without any texture or preferential orientation. After normalization, a comparison between the integrated intensity of each peak, i.e., the area under the peak rather than the height, and its corresponding intensity in the JCPDS file indicates that the NiSi film on the Si(001) substrate does not have a random, powder-like distribution of grain orientations. Rather there is a strong tendency for the grains to align in a certain way with respect to the substrate. Indeed, a thorough pole-figure analysis, a specially arranged diffractometry that determines the distributions of grain orientations, has revealed that the NiSi film is actually strongly "textured" (Figure 10.19) [142]. However, this texture is different from the commonly understood kinds. It is an entirely new kind with an off-normal fiber axis leading to a "one-dimensional" epitaxy termed "axiotaxy"; details about it are found in [22,142].

As the new kind of grain alignment is introduced, it is necessary to recall the conventional classification of texture in thin films that usually comprise a giant number of grains (crystallites). In the literature, texture in thin films has been classified into three different categories: random, fiber, and in-plane texture [191,192]. For the case of random texture, no restriction is imposed on grain orientation and all the peaks are detected with intensities after normalization identical to those in the corresponding JCPDS file. For fiber texture, a specific crystal plane of the crystallites is parallel to the sample surface, leaving a single degree of freedom to rotate the grains around the fiber axis when positioning them on the substrate. This single degree of freedom results in diffraction peaks detected with intensities significantly different from

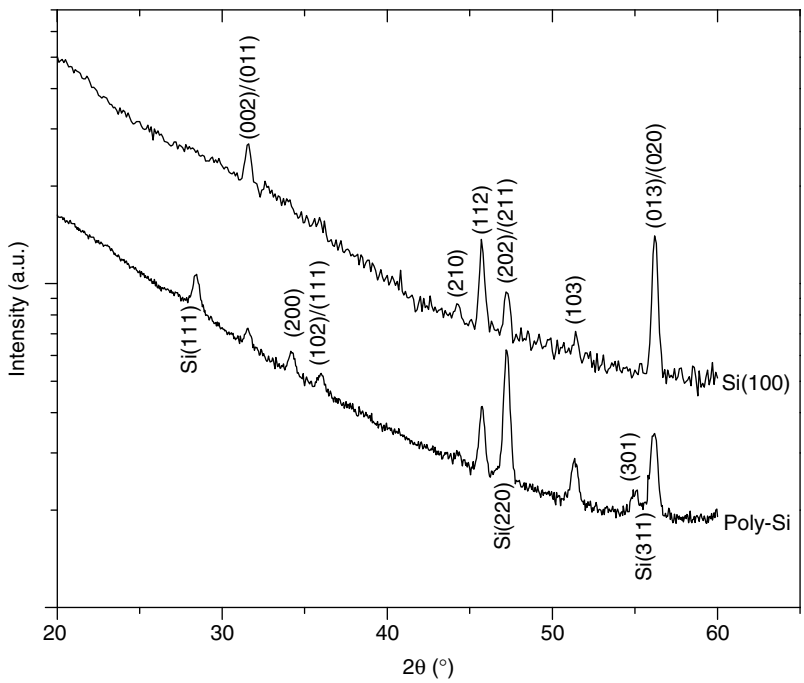


FIGURE 10.18 θ–2θ XRD measurements for NiSi films on Si(001) and poly-Si substrates. (Reproduced from Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004. With permission.)

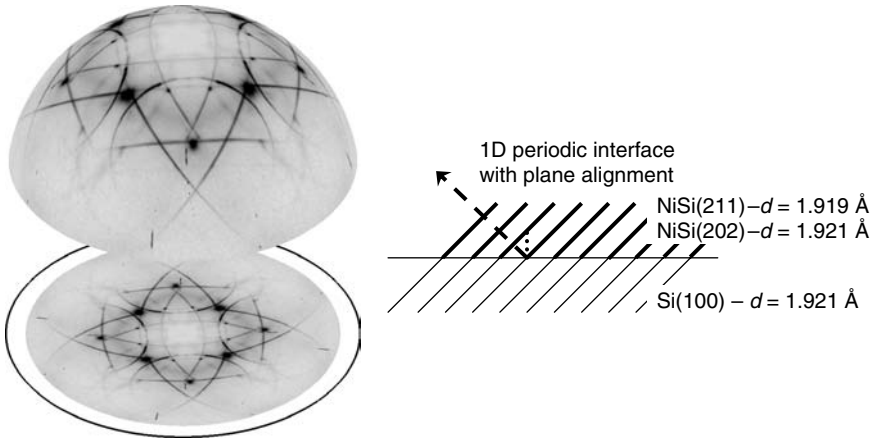


FIGURE 10.19 Pole figure measurement [(112) planes] of a 60 nm NiSi film showing the sharp lines related to axiotaxy. The schematic on the right describes the alignment of planes of same d spacing, necessary to obtain the sharp lines on the pole figure. (Adapted from Lavoie, C., Detavernier, C., and Besser, P., *Nickel Silicide Technology*, IEEE, London, 2004.)

the powder sample counterpart. For in-plane alignment, the orientation of the grain is completely determined by the substrate, leading to a diffraction pattern with a much reduced number of peaks detected with greatly increased intensities. Axiotaxy is then a fourth kind of texture and is the result of plane alignment through the interface as schematically shown in Figure 10.19. The (202) and (211) planes of NiSi grains have very similar *d* spacings as the (220) planes of the Si substrates. When these planes align, it is clear that periodicity develops in one direction. How such an axiotaxy would influence the integrity of a grown NiSi film including the morphological stability is not clear at this time. However, we point out that the low interfacial energy for axiotaxial grains is independent of developing roughness at the interface.

10.4.7 $Si_{1-x}Ge_x$ Devices

Downscaling has evolved into a situation where combining material solutions and architecture innovation with dimensional miniaturization becomes crucial for enhanced device and circuit performance [193,194]. Thus, Ge is incorporated to form $Si_{1-x}Ge_x$ alloys in the key locations of a transistor for various purposes: in the gate for tailoring the work function [195,196]; in or below the channel for boosting the carrier mobility [197–200]; and in the source/drain (S/D) regions for increasing conductivity and reducing contact resistivity [201–206]. This last approach has recently also contributed to an additional advantage in boosting the device performance: epitaxy of $Si_{1-x}Ge_x$ either as a refill in a recess etched S/D [207] or as a raised S/D [208] of a deep sub-100 nm MOSFET is used to generate compressive stresses on the channel thereby enhancing the hole mobility. These different efforts are schematically summarized in Figure 10.20. Contact formation relying on the salicide process will concern interaction between a selected metal and the $Si_{1-x}Ge_x$ in these regions. When $Si_{1-x}Ge_x$ is purposely used in the terminal regions either for work-function adjustment, Figure 10.20a, or for reduction of source/drain resistance, Figure 10.20b, the need for consideration of a metal-$Si_{1-x}Ge_x$ interaction is obvious referring to the salicide process reviewed in Figure 10.2. Current process technology for the fabrication of channel structures either with a buried strained $Si_{1-x}Ge_x$ channel, Figure 10.20c, or with a surface strained Si channel grown on a $Si_{1x}Ge_x$ virtual substrate, Figure 10.20d, inevitably leads to identical Si/$Si_{1-x}Ge_x$ layer structures present in the source and drain regions. Since the Si layer in both cases is typically below 10 nm thickness, the silicidation front will undoubtedly reach the underneath

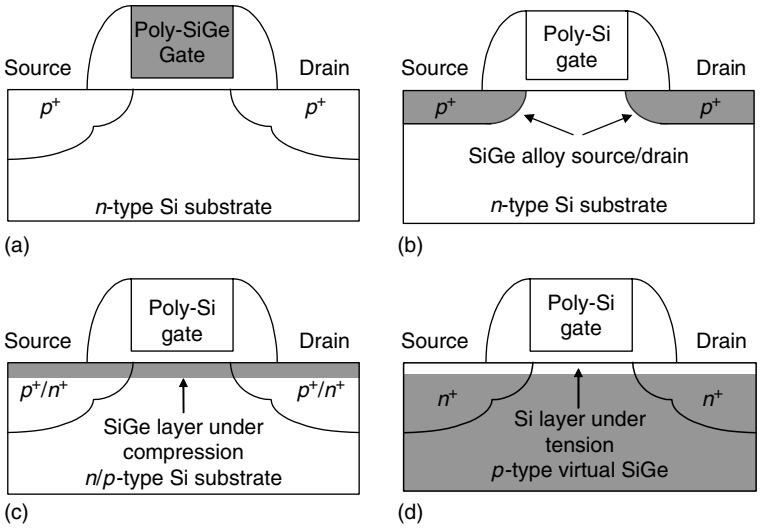


FIGURE 10.20 Incorporation of $Si_{1-x}Ge_x$ in a modern MOSFET: (a) as a replacement of poly-Si in the gate, (b) as a low-resistivity source/drain extension as well as a source of mechanical force compressing the channel region from the two sides, (c) as an epitaxially grown, strained $Si_{1-x}Ge_x$ channel buried under a thin Si cap, and (d) as a (single-crystal) virtual substrate for the growth of a strained Si channel atop.

$Si_{1-x}Ge_x$. Hence, the consideration of a metal-$Si_{1-x}Ge_x$ interaction is also relevant in these two device structures.

10.4.7.1 Choice of Ni, Advantages in Phase Stability and Low Contact Resistivity

In the presence of Ge, contact formation using Ni is especially attractive and viable. A rapidly increasing number of publications on the reaction of Ni with $Si_{1-x}Ge_x$ during the past few years can be found in the literature. Recent reviews about the Ni–$Si_{1-x}Ge_x$ interaction are available [209–211] and here only a brief overview is provided.

The choice of Ni over other metals such as Ti and Co for contact formation in $Si_{1-x}Ge_x$ devices is based on two remarkable advantages [209–236]: (1) formation of the ternary solid solution $NiSi_{1-u}Ge_u$ with a low resistivity below 500°C and (2) increased formation temperature of the high-resistivity phase $NiSi_2$ above 850°C, through an identical mechanism as for the enhanced phase stability of NiSi in the presence of Pt or Pd discussed above. NiSi and NiGe form a ternary alloy with a complete mutual miscibility [212] while $NiGe_2$ does not form as a stable phase under ordinary formation conditions [237]. Calculations of the ternary phase diagram for the Ni–Si–Ge system can be found in [209–213]. Hereafter, u in $NiSi_{1-u}Ge_u$ can be different from x in $Si_{1-x}Ge_x$ because of out-diffusion of Ge from the germanosilicide dictated by thermodynamics [209,212,213]. NiSi has a well-documented low resistivity, see Table 10.3. The resistivity of NiGe thin films is also very low at 15–19 μΩ cm [238]. In spite of alloy scattering, the monogermanosilicide $NiSi_{1-u}Ge_u$ films typically have a resistivity around 20 μΩ cm [147,212].

For Ti, the formation temperature of the low-resistivity C54-$Ti(Si_{1-u}Ge_u)_2$ is too high (above 700°C) to be compatible with the stability requirements of a strained Si or SiGe. Precipitation of Ge from the formed $Ti(Si_{1-u}Ge_u)_2$, also dictated by thermodynamics [239] occurs below 700°C [240,241]. If a poly-Si gate is in use, the high resistance in fine-line discussed earlier remains a serious handicap for the use of Ti. The addition of Ge actually worsens the morphological stability: agglomeration of the $Ti(Si_{1-u}Ge_u)_2$ films on a $polySi_{1x}Ge_x$ substrate occurs around 750°C [77]. For Co, the presence of Ge increases the formation of the low-resistivity $CoSi_2$ by 250°C [242,243]. That CoSi and CoGe are mutually miscible

forming a ternary monogermanosilicide while $CoSi_2$ and $CoGe_2$ do not even share the same crystallographic structure leads to a decreased Δg and increased ΔG^* for the formation of the disilicide–digermanide from the monogermanosilicide [60].

In view of its ability to form the low-resistivity NiSi in small dimension areas, the use of Ni-based salicide process apparently becomes the most viable choice for ultra-scaled devices. This is also true in the presence of Ge. The low contact resistivity around $10^{-8}\ \Omega\,cm^2$ or $1\ \Omega\,\mu m^2$, for $NiSi_{1-u}Ge_u$ contacts achieved by means of above-solubility doping for both conduction types [201–204], boron for *p*-type $Si_{1-x}Ge_x$ and phosphorous for *n*-type $Si_{1-x}Ge_x$, also greatly boosts the confidence with Ni. The above-solubility doping is realized by in situ doping during epitaxy of $Si_{1-x}Ge_x$ in a chemical vapor deposition process. That a much higher concentration of boron than thermodynamically allowed can be incorporated in the otherwise compressively strained, epitaxially grown $Si_{1-x}Ge_x$ on Si(100) is believed to be due to size compensation [244]. The low-temperature formation of $NiSi_{1-u}Ge_u$ is apparently a favorable factor for maintaining the non-equilibrium state with such heavily doped $Si_{1-x}Ge_x$.

10.4.7.2 Challenges with Morphological Stability and Progress

The main shortcoming of $NiSi_{1-u}Ge_u$ is still related to the poor morphological stability; actually the presence of Ge just worsens the situation because of a lowered melting point of the ternary alloy $NiSi_{1-u}Ge_u$ compared to that of NiSi. Typically, degradation of $NiSi_{1-u}Ge_u$ films on $Si_{1-x}Ge_x$ occurs at a temperature 100°C–200°C lower than that for NiSi films on Si depending on the fraction of Ge (value of *x*) as well as the substrate type, polycrystalline or single crystal. The mechanisms responsible for the morphological degradation of $NiSi_{1-u}Ge_u$ films on $Si_{1-x}Ge_x$ substrates are identical to those for the NiSi films on Si: grain growth in the underlying $Si_{1-x}Ge_x$ for poly-$Si_{1-x}Ge_x$ substrates and grain grooving in the $NiSi_{1-u}Ge_u$ films for sc-$Si_{1-x}Ge_x$ substrates. Both processes, grain growth and grain grooving, are again driven by a minimization of free energy of the system in question.

In addition to grain growth in poly-$Si_{1-x}Ge_x$ and grain grooving in $NiSi_{1-u}Ge_u$, two additional effects specific for the systems of $NiSi_{1-u}Ge_u/Si_{1-x}Ge_x$ have been identified [211,222]: (1) the tendency to form NiSi at the expense of NiGe and (2) the mechanical strain in the underlying $Si_{1-x}Ge_x$. The latter mainly concerns the system with $NiSi_{1-u}Ge_u$ on sc-$Si_{1-x}Ge_x$. Through their influence on the atomic compositions at the $NiSi_{1-u}Ge_u/Si_{1-x}Ge_x$ and thus the interface energies, both effects contribute to the morphological instability of the systems. The first effect that is present in both poly- and sc-$Si_{1-x}Ge_x$ systems, leads to out-diffusion of Ge and segregation of Ge-rich $Si_{1-x}Ge_x$ around the $NiSi_{1-u}Ge_u$ grains. The Ge-rich Si–Ge that precipitates between the $Si_{1-x}Ge_x$ and the $NiSi_{1-u}Ge_u$ originates from thermodynamic equilibrium calculations. The precipitation of this layer at the interface and at grain boundaries surely increases the number of interfaces and does not help in increasing the resistance to agglomeration. Under these conditions and dictated by the equilibrium of surface tensions at the triple junctions between the interface and two grain boundaries (see schematic in Figure 10.21) [245]:

$$\gamma_{gb} = 2\gamma_i \cos\beta \tag{10.3}$$

interface roughening and grain grooving are promoted by the presence of Ge. The result of all these is morphological degradation or agglomeration of the $NiSi_{1-u}Ge_u$ layer at lower temperatures. The influence of Ge on interface morphology has indeed been observed [233]; a few percents Ge segregated at the NiSi/Si interface leads to a considerable roughening and grain grooving at the interface, compared to a NiSi/Si interface without Ge.

The second effect is primarily found in strained $Si_{1-x}Ge_x$ epitaxially grown on sc-Si substrate. Experimentally, a considerably worsened morphological stability of $NiSi_{1-u}Ge_u$ on strained $Si_{1-x}Ge_x$ was observed in comparison with that on relaxed $Si_{1-x}Ge_x$ [222]. If a similar mechanism with the focus on the interface energies should also be operative here, one would need to invoke an argument of a lower interface energy between $NiSi_{1-u}Ge_u$ and $Si_{1-x}Ge_x$ when a higher strain is present in the $Si_{1-x}Ge_x$. On strained $Si_{1-x}Ge_x$, the Si–Ge and Si–Si bonds at the interface are under compression. Removal of some Si or Ge atoms from this region would favor the relief of the compressive stresses, with a consequent

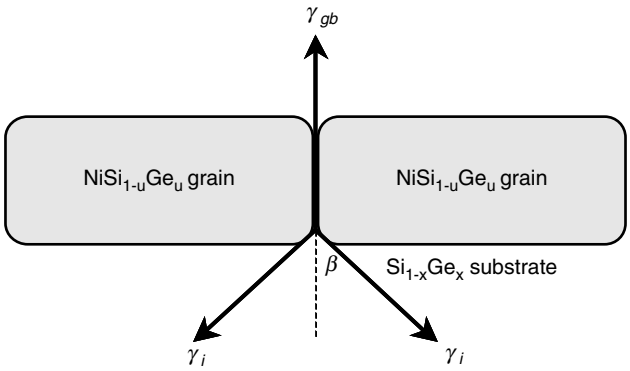


FIGURE 10.21 Schematic force balance at a triple junction involving one interface tension vector and two grain boundary tension vectors.

decrease of the interface energy. Moreover, it is Si atoms that are needed to replace some Ge atoms in the nearby $NiSi_{1-u}Ge_u$ grains [209,213]. A preferential removal of Si atoms leaves the interface at the $Si_{1-x}Ge_x$ side enriched with Ge. This effect, too, would lead to a decrease in interface energy.

To accomplish morphological changes, atomic movements of all constituent atoms are required. The melting point of $NiSi_{1-u}Ge_u$ monogermanosilicide should be lower than NiSi. Hence, the atomic mobility is most likely higher in $NiSi_{1-u}Ge_u$ as compared to NiSi. More importantly, the least mobile species in the $NiSi_{1-u}Ge_u$, Si, and Ge [233,246], are in fact quite mobile leading to an observed substantial inter-diffusion of Si and Ge in $NiSi_{1-u}Ge_u$ already around 550°C [222]. The high mobility of Si and Ge in $NiSi_{1-u}Ge_u$ is quite alike Si in NiSi [92,107] where Ni is also the dominant diffusion species. As a comparison, Si and Ge atoms are practically immobile in $Si_{1-x}Ge_x$/Si below 800°C [247–250]. Therefore, the fast inter-diffusion of Si and Ge inside $NiSi_{1-u}Ge_u$ provides a kinetic pathway for the morphological degradation of the $NiSi_{1-u}Ge_u/Si_{1-x}Ge_x$ system at much reduced temperatures.

Insertion of an ultrathin (1–2 nm) Pt [203,204] or Mo [225] layer at the $Ni/Si_{1-x}Ge_x$ interface or incorporation of C (<1 at%) in the $Si_{1-x}Ge_x$ substrate [220,221,225] has been reported to improve the morphological stability of the resulted $NiSi_{1-u}Ge_u/Si_{1-x}Ge_x$. (Using a Ni–Pt alloy instead of a Pt interlayer has been shown to be effective as well [228].) It can, however, be a major effort to control a system with five elements (Si, Ge, Ni, dopant, and the additive) in any precise sense, despite of the benefit for morphological stability. The improvement due to Pt addition could be attributed to the formation of $PtSi_{1-u}Ge_u$ [251,252], a ternary alloy similar to $NiSi_{1-u}Ge_u$, which is likely to have a higher melting point and thus a lower atomic mobility than $NiSi_{1-u}Ge_u$. Energetically [253], however, PtSi tends to form at the expense of PtGe, similarly to the favored formation of NiSi over NiGe. Much worse, phase separation and $PtGe_2$ formation by destroying the formed $PtSi_{1-u}Ge_u$ have been reported to occur already at 400°C–450°C [251], which is a challenging situation to bear in mind when designing the complex system with Pt, Ni, Si, and Ge. However, why Mo additions could improve the stability is less clear. Although layer inversion occurs leaving the majority of Mo at the surface, segregation of Mo atoms in the $NiSi_{1-u}Ge_u$ grain boundaries is observed [225]. Should the Mo segregation lead to a reduction of the grain boundary energy in the $NiSi_{1-u}Ge_u$ layer relative to the interfacial energy at the $NiSi_{1-u}Ge_u/Si_{1-x}Ge_x$ interface, a more stable $NiSi_{1-u}Ge_u/Si_{1-x}Ge_x$ structure could be anticipated as discussed earlier. One benefit with the use of Mo is the much higher formation temperatures for the silicides and germanides of Mo than for those of Ni. Hence, Mo could be regarded as inert at the temperatures considered, making the formation of Ni-germanosilicide relatively easily controllable. Along this line, W could also be considered. In view of practical difficulties in depositing a 1–2 nm thick Mo interlayer, alloying Mo with Ni could be an interesting alternative. With C in $Si_{1-x}Ge_x$, accumulation of C atoms at

the $NiSi_{1-u}Ge_u/Si_{1-x}Ge_x(C)$ interface was found [221] in addition to a retarded formation of $NiSi_{1-u}Ge_u$. No measurable increase in the specific resistivity of the resulting $NiSi_{1-u}Ge_u$ layers could be detected as a result of a generally low C content used (<1 at%). However, how the interfacial C pile-up would affect the contact integrity electrically remains to be investigated.

Finally, on epitaxial $Si_{1-x}Ge_x$ grown on Si(100) the $NiSi_{1-u}Ge_u$ layers have been found to be strongly textured with (013) planes parallel to the sample surface [234]. When formed at 500°C, the $NiSi_{1-u}Ge_u$ layers, about 50 nm thick, have crystalline grains typically 120–150 nm large. Recent results further show [233] that the thickness as well as the layer sequence, i.e., Si on $Si_{1-x}Ge_x$ vs. $Si_{1-x}Ge_x$ on Si such as those seen in Figure 10.20, strongly influence the fiber texture development of the resulting $NiSi_{1-u}Ge_u$ layer(s). When Ni is alloyed with Pt, the (Ni, Pt)Si layer formed on Si(100) is also preferentially oriented to (013) [254]. These texturing behaviors are different from axiotaxy observed for pure NiSi, without Ge or Pt, on Si(100) as briefly discussed earlier in Section 10.4.6. What the texturing may do to the thin films with respect to the thermal stress has not been fully determined. It is also clear that the CTE of $NiSi_{1-u}Ge_u$ will be anisotropic; texturing in combination with CTE anisotropy may open a window for strain/stress engineering.

10.5 Metal Gate and Schottky Barrier Source-Drain

The use of metal silicides in CMOS technology has witnessed an evolution from WSi_2 and $MoSi_2$ to $TiSi_2$, then to $CoSi_2$ and now to NiSi. In parallel the process technology has shifted from polycide to salicide. The primary purpose has concurrently changed from reducing contact resistance and improving contact reliability, to shunting the high resistivity poly-Si gate, and to improving the contact resistance in the source-drain. In the past 4–5 years, interests in metal silicides have been intensive with two "new" potential application areas: dual work-function metal gate and Schottky barrier source-drain. The former begins from an important discovery of dual effective work function using a single NiSi when fully siliciding the poly-Si gates that are heavily doped to p- and *n*-type on an ordinary SiO_2 gate dielectric [255]. The latter is in fact as a revisit of an almost 40-year old idea to realize ultra-shallow junctions without sacrificing the source-drain series resistance [256], but now with a smart use of complementary silicides with low barrier heights to the channels of opposite polarity [257]. Specifically, $ErSi_{1.7}$ with a Schottky barrier height of 0.3 eV to *n*-type Si [257,258] and PtSi with a Schottky barrier height of 0.21 eV to *p*-type Si [259] were used for the n-channel and p-channel devices, respectively. In these conditions, heavily doped (shallow) p–n junctions are not necessary at the source-drain since the contact is not ohmic anymore. It may be worth mentioning that the concept of using Pt- and Er-silicides as the Schottky gates in complementary Si MESFETs (metal-semiconductor field-effect transistors) was shown almost 20 years ago [260].

10.5.1 Perspectives

The introduction of dual poly-Si gates, *n*-type for n-channel MOSFET and *p*-type for p-channel MOSFET, was a vital step that has facilitated the dimensional downscaling in CMOS technology [261]. The two types of heavily doped poly-Si gate possess distinctly different work functions, one at the conduction band minimum (*n*-type) and one at the valence band maximum (*p*-type) spanning over the entire bandgap of Si. The simultaneous use of the two types of poly-Si gate, yet in two devices of opposite polarity, makes it possible for threshold voltage control at reduced supply voltages. Requiring an identical type of dopant, the doping of the poly-Si gate through ion implantation can be realized in parallel with the doping of the source-drain regions for a MOSFET. The device process has, hence, not been complicated by the use of dual poly-Si gate in CMOS technology, in comparison with the use of a single *n*-type poly-Si gate for both types of MOSFETs. In contrast, the needed buried channel formation in order to reduce the threshold voltage for p-channel MOSFET while using *n*-type poly-Si gate [262] is now omitted by using a *p*-type poly-Si gate.

A critical component in a drastically down-scaled device is the thickness of the gate dielectric. The dimensional downscaling requires that the capacitance equivalent thickness (CET) of the gate dielectric be accordingly reduced in order for the gate to effectively control the channel, without sacrificing the gate leakage due to tunneling. High-permittivity dielectric materials have therefore been investigated aiming at replacing the long-used SiO_2 as the gate dielectric [263]. Since carrier depletion of the poly-Si gate at the interface towards the gate dielectric [264] contributes to thickening of CET, great efforts have concomitantly been initiated on searching for suitable metals with adequate work-functions to replace the poly-Si gates.

There are several advantages with the implementation of Schottky barrier source-drain in a MOSFET. The most important ones are as follows. Firstly, the fabrication process becomes simpler without the involvement of p–n junction formation that is realized through ion implantation and thermal activation at source and drain. In the scaled devices, ion implantation at extremely low energy and activation above the solid solubility limit are needed to attain a low resistivity source-drain extension as well as the ohmic contact with a low resistivity silicide. These steps are not only costly in terms of processing but also require new tools for low-energy implantation and spike annealing. Secondly, since the silicides used for the formation of Schottky contacts are usually those that require low temperature for their formation the process also only requires relatively low temperatures. Typically, a thermal treatment around 500°C can complete the formation [257]. Low temperature processing is particularly attractive for ultra-scaled devices to retain the basic physical and electrical structures such as doping profile in the proximity or within the channel. Low temperature processing has lately also become vital for devices that involve intentionally introduced strain in the channel, as an example, as well as that incorporate high-permittivity dielectrics as the gate dielectric and metals as the gate electrode. Thirdly, the junction can be made much shallower than ordinary p–n junctions. Because the metals used for the formation of Schottky diodes are usually of good conductivity, a very thin layer of such metals is sufficient to provide the needed current without imposing constraints on power consumption. Note that dopants outside the channel cannot be eliminated, as a shallow junction remains crucial for controlling short-channel effects.

The choice of a suitable metal for the formation of the Schottky barrier source-drain in a MOSFET is primarily determined by its work function. (One should however keep in mind that the precise interface structure through film texture, substrate orientation or any impurities at the interface plays a crucial role in influencing the actual Schottky barrier height [265].) A large work function indicates a large barrier height to the *n*-type Si, whereas the opposite is true that a small work function gives rise to a large barrier height to the *p*-type Si [259]. What is desired in a Schottky barrier source-drain MOSFET is a low junction leakage requiring a large Schottky barrier height to the substrate. The metal that possesses a large Schottky barrier height to the substrate then simultaneously yields a low Schottky barrier to the inversion layer that forms the channel. This means an easy carrier injection to the channel and thus a high drain current, a desirable situation.

10.5.2 Fully Silicided Gates: Dual Effective Work-Function

In addition to yielding the desired work-function, metals that can be considered as a metal gate should meet the following requirements:

- Process compatibility with standard CMOS technology in terms of deposition, etching, and cleaning
- Thermal stability and mechanical compatibility with the underlying, delicate gate dielectric films
- Low specific resistivity.

Although many metals or metal systems may meet the physical requirements, those that can be directly integrated into existing process flows should have the best possibility and potential. As mentioned earlier, the industry will try to make the least effort in changing their existing process and process technology. Several interesting concepts along this line, but without the use of metal silicides, include deposition of a

Ti [266] or Mo [267] film followed by ion implantation of nitrogen only into one side of the CMOS devices; deposition and inter-diffusion of multi-layered metal films [268]. (Similar attempts motivated by those original ideas, but use other metal systems, have also been published.) The recently discovered [255] and implemented [267,269,270] FUSI with a single NiSi for both types of MOSFETs has been a very promising approach owing to its compatibility with standard CMOS process. The reminder of this subsection will therefore be focused on this exciting concept.

The discovery with a single NiSi gate for both types of MOSFETs basically implies no fundamental change to the current process technology when replacing the p- and *n*-type heavily doped poly-Si with metallic gates of adequate work functions for p-channel and n-channel MOSFETs, if the corresponding process is controllable and stable. The existing process with ion implantation to simultaneously dope the gate, source, and drain regions for each type of MOSFETs remains unaltered. A major parameter to change could be the thickness of the poly-Si layer; it should be thinner than usual, more or less determined by the source/drain junction depth in order to keep the junction integrity upon NiSi formation. Alternately, the poly-Si thickness could remain large and the full silicidation of the gate could occur at a different stage than the silicidation of the sources and drains. That it is NiSi that gives this unique property with dual effective work function is a gift of the nature, since NiSi is the metal silicide used in extremely scaled devices.

The reason behind the duality in effective work function has been attributed to the formation of a dipole layer between the NiSi and the underlying SiO_2. This dipole layer is primarily composed of dopant atoms originally incorporated in the poly-Si layers. The effects of B, P or As have been quite consistently reproduced from different research groups [255,267,269]. When Al or Sb is implanted instead of B or P and As, the diffusion of Al atoms down to the SiO_2 interface leads to the formation of Al_2O_3 whereas an interfacial layer of Sb with a metallic nature is found [271,272]. How these two dopants would influence the effective work function in actual MOSFETs remains to be investigated.

The duality of effective work function has also been studied for $CoSi_2$ [269,273], PtSi [269], and TiSi [274] (but it is unclear what the Ti-silicide phase actually was). Experimental attempts have further been reported with an incorporation of alloying elements such as Co [269,275] or Pt [276] into the Ni–Si system. All such attempts seem to have led to similar effects in causing the duality, with the assistance of the dopant atoms (B, P, and As), although the precise amplitude of the difference in the two effective work functions differs from one material system to another. Despite of all those attempts, the simplicity in processing, the stability with the underlying SiO_2, and the ability in giving rise to the desired dual effective work function spanning almost over the entire badgap of Si have made NiSi the most studied, and also most promising, fully silicided metal gate candidate.

It should be noted that the observed dual effective work function has been reported for NiSi formed on SiO_2. So is the case for the other silicides mentioned above. Such a duality is, most likely, difficult to extend to and apply for gate stacks where the gate dielectric is changed from SiO_2 to a high-permittivity dielectric such as HfO_2. A thorough analysis of a large compilation of metal-dielectric systems studied for metal oxide semiconductor (MOS) devices has led to the conclusion [277,278] that an interface dipole between metallic gate electrode and gate dielectric, caused by metal-induced gap states [279], plays a critical role in determining the effective work function. The interface dipole model has further been employed to account for thermal instability of metal gate work functions [280]. Hence, even if a duality would still be observable for a metal silicide when switching from SiO_2 to a specific high-permittivity dielectric, the amplitude of the difference in the two effective work functions could hardly be comparable. Consequently, the study and search for metal systems are now combined with the selected high-permittivity dielectrics.

10.5.3 Silicide Schottky Barrier Source-Drain

In devices with very short channel lengths, contact resistance becomes a significant part of the device overall resistance, leading to a search for the minimization of the contact resistance and therefore to

TABLE 10.4 Schottky Barrier Height Φ_B (in eV) for Several Metal Silicides Potentially Interesting for Schottky Barrier Source-Drain Applications

Silicide	Φ Bn (eV)	Φ Bp (eV)	Remarks
$OsSi_{1.8}$	0.85	—	Growth by Si diffusion
LrSi	0.93	—	Growth by Si diffusion
lr_2Si_3	0.85	—	Growth by Si diffusion
Pt_2Si	0.85	—	Growth by Si and Pt diffusion
PtSi	0.88	0.21	Growth by Si and Pt diffusion
$lrSi_3$	—	0.94	Growth by Si diffusion
YSi_2	0.39 ± 0.03	—	Growth by Si diffusion
$GdSi_{2-x}$	0.37 ± 0.02	0.71 ± 0.03	Growth by Si diffusion
$TbSi_{2-x}$	0.38	0.74	Growth by Si diffusion
$DySi_{2-x}$	0.37 ± 0.02	0.73 ± 0.03	Growth by Si diffusion
$HoSi_{2-x}$	0.37 ± 0.02	—	Growth by Si diffusion
$ErSi_{2-x}$	0.39 ± 0.02	0.70 ± 0.02	Growth by Si diffusion
$YbSi_2$	0.27	0.85	Growth by Si diffusion

Φ_{Bn} on *n*-type Si and Φ_{Bp} on *p*-type Si. Dominant diffusion species for the growth of the various silicides also specified.
Source: From Gas P. and d'Heurle, F. M., *Properties of Metal Silicides*, INSPEC, London, 1995; Derrien, J., *Properties of Metal Silicides*, INSPEC, London, 1995; Tung, R. T., *Mater. Sci. Eng.: R: Reports* 35, 138, 2001; Tu, K. N., Thompson, R. D., and Tsaur, B. Y., *Appl. Phys. Lett.* 38, 626, 1981; Norde, H., de Sousa Pires, J., d'Heurle, F., Pesavento, F., Petersson, S., and Tove, P. A., *Appl. Phys. Lett.* 38, 865, 1981; Zhu, S., Chen, J., Li, M.-F., Lee, S. J., Singh, J., Zhu, C. X., Du, A., Tung, C. H., Chin, A., and Kwong, D. L., *IEEE Electron Device Lett.* 25, 565, 2004.

contacts with sufficiently low barrier heights. Replacing deep *n–p* or *p–n* junctions at source and drain with Schottky contacts can in principle be achieved using pure metals. However, pure metals are ruled out since during processing they would react with Si to form silicides. Therefore, the discussion below is focused on these compounds. A brief survey of the Schottky barrier heights in the available handbooks [53,124,281] indicates that only a limited number of metal silicides display a sufficiently large value of Schottky barrier height that is potentially attractive for either *n*- or *p*-channel MOSFETs. These silicides are listed in Table 10.4, the sources are extended to some original articles [282–284]. Also given, in the last column, is the dominant diffusion species during the growth of the various silicides [125].

An adequate work function or desired Schottky barrier height is not sufficient for the selection of a silicide for device applications. As an example, $IrSi_3$ with its largest Schottky barrier height to *p*-type Si, 0.94 eV is formed by a nucleation controlled process from its preceding phase Ir_2Si_3 [57]. This formation characteristic not only indicates a rough interface between the resulting silicide layer and the Si substrate, but also leads to a much increased temperature for formation. As a result of its well-behaving growth controlled by diffusion, PtSi has so far been the mostly studied silicide for Schottky barrier source-drain applications in p-channel MOSFETs in the literature [284–286], even when SiGe is present in the source-drain areas [287]. PtSi has not only been successfully incorporated in Schottky barrier source-drain MOSFETs fabricated in 10 nm thick ultrathin SOI [257,288], but also shown good potential in Schottky barrier source-drain FinFETs [289].

For n-channel MOSFETs [284,286,290], the success with $ErSi_{1.7}$ [257] has seen several followers [284,286,290]. In particular, the $YbSi_{2-x}$ has recently been singled out [284] as the most promising candidate since this silicide phase is found to have the highest barrier to *p*-type Si and thus the lowest barrier to the *n*-type channel. As a result, the use of this silicide has been shown to give rise to a lower junction leakage and a higher drain current than the other rare earth metal silicides in Table 10.4. The current ratio of the on- to off-state is 10^6. In Table 10.4, $YbSi_{2-x}$ along with PtSi is singled out as the most promising candidates for the Schottky barrier source-drain applications. Unlike the transition metals such as Ti, Co, Ni, and Pt, the rare-earth metals used for n-channel MOSFETs are highly reactive. Oxidation of rare-earth metal films can readily occur prior to the silicide formation. A proper surface protection is needed. Usually this can be achieved through deposition of the metals in an ultrahigh

vacuum system followed by deposition of a suitable cap layer without breaking the vacuum. The cap should be in a material affecting neither the silicide formation nor the subsequent selective removal of the unreacted metal on the surrounding isolation.

Several additional concerns on a reliable formation of metal silicides as Schottky barrier junctions are discussed as follows. A large mismatch in the thermal expansion coefficients is present between silicides and Si and SiO_2. A physical discontinuation, upon cooling, at the interface between the silicide front and the Si channel is a disaster for the device performance. A low process temperature is thus important to minimize the thermal mismatch. Furthermore, all the potentially interesting silicides, except Pt_2Si and PtSi, are formed with Si as the dominant diffusion species as seen in Table 10.4; in Pt_2Si and PtSi, Si, and Pt seem to be comparably mobile [125,281]. Now there is a delicate situation to consider. In an ordinary MOSFET structure prior to the salicide process, spacers are present on the two sides of a gate, as seen in Figure 10.2. Assuming the spacers are retained for the Schottky barrier devices, the source-drain regions including the extensions under the spacers are lightly doped (or undoped). In order for the Schottky junctions to form undisturbed, a large series resistance is anticipated if the Si under the spacers remains intact without being silicided. A low on-current is thus expected. A comparative study has shown the importance of controlling the electrical conductance under the spacers with PtSi as the Schottky barrier source-drain [289]. This situation can become particularly serious if Si is the dominant diffusion species and the metal does not move during the silicide formation. The silicide will, in this case, be confined outside the feet of the spacers, strictly in the source-drain terminals. Minimizing the width of the spacers is effective in reducing the high series resistance [289], on the condition that the electrical isolation between the terminals is not tolerated. But as long as the spacers are present, the problem is difficult to completely eliminate and will become particularly problematic for extremely down-scaled devices. On the other hand, it is risky, and thus not desirable, to let the silicide front erode the Si under the spacers and then continue to move uncontrolled into the channel region. Although a low series resistance is attained, this uncontrolled silicide growth is particularly harmful for scaled devices with a gate length below 50 nm. This situation can occur if the metal is the dominant diffusion species. It can become especially pronounced with a limited Si supply; a drastic lateral growth of NiSi into the channel region has been observed when formed on ultra-thin SOI substrates [291–293]. Apparently, the design of the spacers needs to take into account the growth mechanism and formation kinetics of the silicides.

One likely device structure for a future MOSFET will feature a nanowire channel with a cross-sectional dimension below 10 nm [193]. With a wrapped gate as shown in Figure 10.22, such a device has theoretically been shown to yield the best electrostatics for short-channel effects for a given channel length, in comparison with single-gate (planar), double-gate, or tri-gate MOSFETs [294]. As regard to the contact formation in the source and drain regions, a wrapped contact geometry has been shown, through simulation, to yield the lowest contact resistance to heavily doped source-drain regions in FinFETs [295]. Realizing such a device structure in experiment involves a number of innovations in processing such as selective epitaxy of Si around a Si-nanowire channel prior to the wrapped silicide contact formation [296]. For nanowire devices with a channel dimension below 10 nm in cross-section,

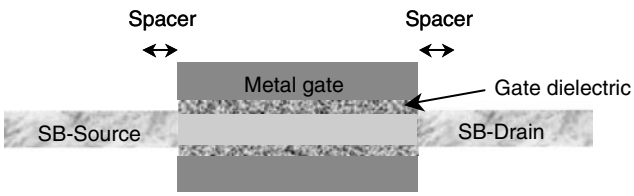


FIGURE 10.22 Schematic cross-section (not to scale) of a nanowire MOSFET with a wrapped gate stack and a Schottky barrier source-drain (SB-Source and SB-Drain).

the extremely small volume of Si can lead to random discrete dopants in the source-drain regions even if the doping level reaches the solid solubility. This randomness in turn can cause large instability and variations of the contact resistance and thus in the current. A Schottky barrier source-drain can be an attractive alternative, avoiding the difficulty with dopants, cf. Figure 10.22. Apart from the process challenges discussed earlier, i.e., diffusion and spacers, a major technological challenge with silicides as Schottky barrier source-drain concerns the quite large Schottky barrier heights (>0.1 eV) of the known and available silicides to the Si channels of either type. Finally, the silicide in direct contact with the Si nanowire-channel will also be in a nanowire form. For a Si nanowire with such an extreme dimension, it will be a single silicide grain that is in contact with the channel. This situation is different from planar MOSFETs either on bulk Si or in UTB SOI where a string of silicide grains will simultaneously be contacting the channel along the transistor width. Potential variations in the actual Schottky barrier height from device to device could now arise for nanowire MOSFETs, mainly due to two factors: (1) the variation of metal work function with its crystallographic surface [297] and (2) the critical influence of the interface dipole at the silicide-Si interface on the actual Schottky barrier height [265]. In these dimensions, the variations expected from one device to the next are anticipated to be very significant and much remains to be investigated and clarified by further research.

10.6 Summary

This chapter attempted to cover the history and recent developments of metal-silicides in CMOS technology. The SALICIDE process was described with its evolution from $TiSi_2$, $CoSi_2$ and NiSi. These silicides extensively used for contact in the microelectronic industry were sequentially described with a particular emphasis on the current NiSi process and properties. The $TiSi_2$ contacts were limited by the lack of transformation to the low resistivity C54 phase in dimensions smaller than 0.2–0.3 μm. The $CoSi_2$ contacts did not show this type of drawback but were limited by void formation in very narrow poly-Si lines (<50 nm). The introduction of SiGe substrates to bolster the mobility of carriers is also found to be incompatible with the standard Co process. The introduction of Ni silicide, while showing numerous advantages, leads to many constraints on the overall process, as the temperature following silicidation needs to be drastically reduced to avoid the thermal degradation of this very sensitive material. Although studied for many decades, the increased focus on this low resistivity material pointed out unexpected properties such as the presence of multiple metal-rich phases in the phase formation sequence and the important new type of texture (axiotaxy) in which planes of "equal" spacing in the NiSi and the Si align through the interface. The known large anisotropy in thermal expansion was also emphasized. In the last section of the chapter, alternative silicides and possible use of silicides for FUSI were also presented. It is difficult to present an accurate overview of these last subjects, as the industry is currently moving extremely rapidly. As the stakes are incredibly high, the amount of work in progress is astounding. A simple review of the latest IEDM and very large scale integration (VLSI) meetings show numerous results that already modify or complete some of the thoughts presented in this last section.

Acknowledgments

The authors thank Cyril Cabral, Christophe Detavernier, Roy Carruthers, Robert Purtell, Cedrik Coia, Simon Gaudet, Jean Jordan-Sweet, Paul Besser, Jorge Kittl, Aliette Mouroux, Johan Seger, Tobias Jarmar, Zhen Zhang, Per-Erik Hellström, and Ulf Smith for contributions to the experiments and stimulating discussions. Multiple figures, ideas, and summarized descriptions were reproduced with permissions from Refs. 80 and 22. The synchrotron XRD experiments were conducted under DOE contract DE-AC02-76CH-00016. S.-L. Zhang thanks the Swedish Foundation for Strategic Research (SSF), the Swedish Science Council (VR), and the Swedish National Council for Engineering Research (TFR) for financial support.

References

1. Kahng, D., and M. P. Lepselter. *Bell Syst. Technol. J.* 44 (1965): 1525.
2. Sinha, A. K. *J. Electrochem. Soc.* 120 (1973): 1771.
3. Kircher, C. J. *J. Appl. Phys.* 47 (1976): 5394.
4. Kircher, C. J. *J. Electrochem. Soc.* 14 (1971): 507.
5. Sze, S. M. *Physics of Semiconductor Devices*, 303–4. NY: Wiley, 1981.
6. Tsai, M. Y., H. H. Chao, L. M. Ephrath, B. L. Crowder, A. Cramer, R. S. Bennett, C. J. Lucchese, and M. R. Wordeman. *J. Electrochem. Soc.* 128 (1981): 2207.
7. Gambino, J. P., and E. G. Colgan. *Mater. Chem. Phys.* 52 (1998): 99.
8. Shibata, T., K. Hieda, M. Sato, M. Konaka, R. L. M. Dang, and H. Iizuka. *IEEE International Electron Devices Meeting 1981—Technical Digest*, 647. Piscataway, NJ; Washington, DC, 1981.
9. Shibata, T., K. Hieda, M. Sato, M. Konaka, R. L. M. Dang, and H. Iizuka. *IEEE Trans. Electron Devices* 29 (1982): 531.
10. Lau, C. K., Y. C. See, D. B. Scott, J. M. Bridges, R. D. Davies, and S. M. Perna. *IEEE International Electron Devices Meeting 1982*, 714. New York, NY; San Francisco, CA, 1982.
11. Osburn, C. M., M. Y. Tsai, S. Roberts, C. J. Lucchese, and C. Y. Ting. *VLSI Science and Technology*, edited by C. Dell'Oca and W. M. Bullis, 213–23. Pennington: The Electrochemical Society, 1982.
12. d'Heurle, F. M. *J. Electron. Mater.* 27 (1998): 1138.
13. Handbook of Binary Alloy Phase Diagrams (CD version 1.0), ASM International.
14. Jiang, Y.-L., G.-P. Ru, W. Huang, X.-P. Qu, B.-Z. Li, A. Agarwal, G. Cai, J. Poate, C. Detavernier, and R. L. Van Meirhaeghe. *Semicond. Sci. Technol.* 20 (2005): 716.
15. Ru, G.-P., Y.-L. Jiang, X.-P. Qu, and B.-Z. Li. *2004 7th International Conference on Solid-State and Integrated Circuits Technology Proceedings* (IEEE Cat. No. 04EX862), 451. Beijing, China: IEEE, 2005.
16. Mo, H., P. Bonfanti, B. Zhu, D. Gao, H. Wu, J. Chen, H.-Z. Wu, Y.-L. Jiang, G.-P. Ru, and F. Chen. *2004 7th International Conference on Solid-State and Integrated Circuits Technology Proceedings* (IEEE Cat. No. 04EX862), 464. Beijing, China: IEEE, 2005.
17. Foggiato, J., W. S. Yoo, M. Ouaknine, T. Murakami, and T. Fukada. *Mater. Sci. Eng. B* (*Solid-State Materials for Advanced Technology*) 114–5 (2004): 56.
18. Lauwers, A., J. A. Kittl, M. van Dal, O. Chamirian, R. Lindsay, M. de Potter, C. Demeurisse., et al. *Microelectron. Eng.* 76 (2004): 303.
19. Lavoie, C., C. Cabral Jr., L. A. Clevenger, J. M. E. Harper, J. Jordan-Sweet, K. L. Saenger, and F. Doany. *Mater. Res. Soc. Symp. Proc.* 406 (1996): 163.
20. Stephenson, G. B., K. F. Ludwig, J. L. B. S.Jordan-Sweet, J. Mainville, Y. S. Wang, and M. Sutton. *Rev. Sci. Instrum.* 60 (1989): 1537.
21. Zhang, S.-L., C. Lavoie, C. Cabral Jr., J. M. E. Harper, F. M. d'Heurle, and J. Jordan-Sweet. *J. Appl. Phys.* 85 (1999): 2617.
22. Lavoie, C., C. Detavernier, and P. Besser. "Nickel Silicide Technology." In *Silicide Technology for Integrated Circuits*. London: IEEE, 2004.
23. Nygren, S., and F. d'Heurle. *Diffusion Defect Data—Solid State Data Part B* (*Solid State Phenomena*) 23–24 (1992): 81.
24. Yan, Z. H., M. Oehring, and R. Bormann. *J. Appl. Phys.* 72 (1992): 2478.
25. Mann, R. W., L. A. Clevenger, P. D. Agnello, and F. R. White. *IBM J. Res. Dev.* 39 (1995): 403.
26. Mann, R. W., and L. A. Clevenger. *J. Electrochem. Soc.* 141 (1994): 1347.
27. Ozcan, A. S., K. F. Ludwig Jr., C. Lavoie, C. Cabral Jr., J. M. E. Harper, and R. M. Bradley. *J. Appl. Phys.* 92 (2002): 5189.
28. Svilan, V., K. P. Rodbell, L. A. Clevenger, C. Cabral Jr., R. A. Roy, C. Lavoie, J. Jordan-Sweet, and J. M. E. Harper. *J. Electron. Mater.* 26 (1997): 1090.
29. Kittl, J. A., Q. Z. Hong, M. Rodder, and T. Breedijk. *IEEE Electron Device Lett.* 19 (1998): 151.
30. Tung, R. T. *Mater. Res. Soc. Symp. Proc.* 402 (1996): 101.
31. Chang, S. M., H. Y. Yang, H. Y. Huang, and L. J. Chen. *Res. Soc. Symp. Proc.* 564 (1999): 65.

32. Mann, R. W., G. L. Miles, T. A. Knotts, D. W. Rakowski, L. A. Clevenger, J. M. E. Harper, F. M. D'Heurle, and C. Cabral. *Appl. Phys. Lett.* 67 (1995): 3729.
33. Mann, R. W., L. A. Clevenger, G. L. Miles, J. M. E. Harper, C. Cabral Jr., F. M. d'Heurle, T. A. Knotts, and D. W. Rakowski. *Mater. Res. Soc. Symp. Proc.* 402 (1996): 95.
34. Mouroux, A., S.-L. Zhang, W. Kaplan, S. Nygren, M. Östling, and C. S. Petersson. *Appl. Phys. Lett.* 69 (1996): 975.
35. (a) Mouroux, A., S. -L. Zhang, and C. S. Petersson. *Phys. Rev. B* 56 (1997): 10614. (This work concerns Ta-interlayer.); (b) Åberg, J., S. Persson, P. -E. Hellberg, S. -L. Zhang, U. Smith, F. Ericson, M. Engström, and W. Kaplan. *J. Appl. Phys.* 90 (2001): 2380. (This work concerns Nb-interlayer.)
36. Cabral, C., L. A. Clevenger Jr., J. M. E. Harper, F. M. d'Heurle, R. A. Roy, K. L. Saenger, G. L. Miles, and R. W. Mann. *J. Mater. Res.* 12 (1997): 304.
37. Cabral, C. Jr., L. A. Clevenger, J. M. E. Harper, F. M. d'Heurle, R. A. Roy, C. Lavoie, K. L. Saenger, G. L. Miles, R. W. Mann, and J. S. Nakos. *Appl. Phys. Lett.* 71 (1997): 3531.
38. Kappius, L., and R. T. Tung. *Mater. Res. Soc. Symp. Proc.* 611 (2001): 821.
39. Zhang, S.-L., F. M. d'Heurle, C. Lavoie, C. Cabral Jr., and J. M. E. Harper. *Appl. Phys. Lett.* 73 (1998): 312.
40. Harper, J. M. E., C. Cabral, and C. Lavoie. *Ann. Rev. Mater. Sci.* 30 (2000): 523.
41. La Via, F., F. Mammoliti, and M. G. Grimaldi. *Appl. Phys. Lett.* 85 (2004): 5577.
42. Chong, Y., K. L. Pey, A. T. S. Wee, A. See, Z. X. Shen, C.-H. Tung, R. Gopalakrishnan, and Y. F. Lu. *J. Electron. Mater.* 30 (2001): 1549.
43. Larciprete, R., M. Danailov, A. Barinov, L. Gregoratti, and M. Kiskinova. *J. Appl. Phys.* 90 (2001): 4361.
44. Chen, S. Y., Z. X. Shen, S. Y. Xu, C. K. Ong, A. K. See, and L. H. Chan. *J. Electrochem. Soc.* 149 (2002): 609.
45. van Gurp, G. J., and C. Langereis. *J. Appl. Phys.* 46 (1975): 4301.
46. Cabral, C. Jr., K. Barmak, J. Gupta, L. A. Clevenger, B. Arcot, D. A. Smith, and J. M. E. Harper. *J. Vac. Sci. Technol.* A11 (1993): 1435.
47. Lavoie, C., C. Cabral, F. M. d'Heurle, and J. M. E. Harper. *Defect Diffus. Forum* 194–199 (2001): 1477.
48. Lavoie, C., C. Cabral Jr., F. M. d'Heurle, J. L. Jordan-Sweet, and J. M. E. Harper. *J. Electron. Mater.* 31 (2002): 597.
49. Goto, K., A. Fushida, J. Watanabe, T. Sukegawa, Y. Tada, T. Nakamura, T. Yamazaki, and T. Sugii. *IEEE Trans. Electron Devices* 46 (1999): 117.
50. Tung, R. T. *Jpn. J. Appl. Phys.* 36 (1997): 1650.
51. Tung, R. T. *Appl. Phys. Lett.* 68 (1996): 3461.
52. Maex, K., A. Lauwers, P. Besser, E. Kondoh, M. de Potter, and A. Steegen. *IEEE Trans. Electron Devices* 46 (1999): 1545.
53. Maex, K., and M. von Rossum, Eds. *Properties of Metal Silicides*. London: INSPEC, 1995.
54. d'Heurle, F. M. *Mater. Res. Soc. Symp. Proc.* 25 (1984): 3.
55. Anderson, R., J. Baglin, J. Dempsey, W. Hammer, F. d'Heurle, and S. Petersson. *Appl. Phys. Lett.* 35 (1979): 285.
56. d'Heurle, F. M., and C. S. Petersson. *Thin Solid Films* 128 (1985): 283.
57. d'Heurle, F. M. *J. Mater. Res.* 3 (1988): 167.
58. Agnello, P. D., S. Brodsky, E. Crabbe, E. Nowak, J. Lasky, and B. Davari. *Electrochem. Soc. Symp. Proc.* 99-100 (1999): 217.
59. Madelung, O. *Semiconductors Basic Data*. New York: Springer, 1996.
60. Detavernier, C., T. R. L. Van Meirhaeghe, F. Cardon, and K. Maex. *Thin Solid Films* 384 (2001): 243.
61. Detavernier, C., R. L. Van Meirhaeghe, F. Cardon, and K. Maex. *Phys. Rev. B* 62 (2000): 12045.
62. d'Heurle, F. M., D. D. Anfiteatro, V. R. Deline, and T. G. Finstad. *Thin Solid Films* 128 (1985): 107.
63. Bourret, A., F. M. d'Heurle, F. K. Le Goues, and A. Charai. *J. Appl. Phys.* 67 (1990): 241.
64. Zirinsky, S., W. Hammer, F. d'Heurle, and J. Baglin. *Appl. Phys. Lett.* 33 (1978): 76.
65. Lu, J. P., D. Miles, J. Zhao, A. Gurba, Y. Xu, C. Lin, M. Hewson, et al. *IEEE International Electron Devices Meeting 2002*, 371, 2002.

66. Chung, J. H., J. E. Lee, J. S. Park, J. H. Ku, E. J. Lee, S. H. Park, S. T. Kim, J. L. Nam, and S. I. Lee. *Advanced Metallization Conference—Proceedings*, 495. 2000.
67. Xiang, Q., C. Woo, E. Paton, J. Foster, B. Yu, and M.-R. Lin. "Symposium on VLSI Technology." *Digest of Technical Papers*, 76, 2000.
68. Chau, R., J. Kavalieros, B. Roberds, R. Schenker, D. Lionberger, D. Barlage, B. Doyle, R. Arghavani, A. Murthy, and G. Dewey. *Intern. Electron Devices Meet. Technical Digest. IEDM* (2000): 45.
69. Finstad, T. G., D. D. Anfiteatro, V. R. Deline, F. M. D'Heurle, P. Gas, V. L. Moruzzi, K. Schwarz, and J. Tersoff. *Thin Solid Films* 135 (1986): 229.
70. Lauwers, A., M. de Potter, O. Chamirian, R. Lindsay, C. Demeurisse, C. Vrancken, and K. Maex. *Microelectron. Eng.* 64 (2002): 131.
71. *PDF-2 Database* (JCPDS-International Center for Diffraction Data, 12 Campus Blvd, Newton Square, PA 19073–3273.
72. Villars, P., and L. D. Calvert, Eds. *Pearson's Handbook of Crystallographic Data for Intermetallic Phases*. Metal Park, Ohio: American Society for Metals, 1985.
73. Steegen, A., and K. Maex. *Mater. Sci. Eng. R: Reports* R38 (2002): 1.
74. Iwai, H., T. Ohguro, and S.-I. Ohmi. *Microelectron. Eng.* 60 (2002): 157.
75. Fitzer, E. *Z. Metallkde.* 44 (1953): 462.
76. Wortman, J. J., and R. A. Evans. *J. Appl. Phys.* 36 (1965): 153.
77. Zhang, S.-L., and M. Ostling. *Crit. Rev. Solid State Mater. Sci.* 28 (2003): 129.
78. Zhang, S.-L., and U. Smith. *J. Vac. Sc. Technol.* A22 (2004): 1361.
79. Lavoie, C., R. Purtell, C. Coïa, C. Detavernier, P. Desjardins, J. Jordan-Sweet, C. Cabral, F. M. d'Heurle, and J. M. E. Harper. *Electrochem. Soc. Symp. Proc.* 2002/11 (2002): 455.
80. Lavoie, C., F. M. d'Heurle, C. Detavernier, and C. Cabral. *Microelectron. Eng.* 70 (2003): 144.
81. Gergaud, P., C. Rivero, M. Gailhanou, O. Thomas, B. Froment, and H. Jaouen. *Mater. Sci. Eng. B (Solid-State Materials for Advanced Technology)* B114–115 (2004): 67.
82. Rivero, C., P. Gergaud, M. Gailhanou, O. Thomas, B. Froment, H. Jaouen, and V. Carron. *Appl. Phys. Lett.* 87 (2005): 41904.
83. Beyers, R., and R. Sinclair. *J. Appl. Phys.* 57 (1985): 5240.
84. Tinani, M., A. Mueller, Y. Gao, E. A. Irene, Y. Z. Hu, and S. P. Tay. *J. Vac. Sci. Technol.* B19 (2001): 376.
85. d'Heurle, F., C. S. Petersson, J. E. E. Baglin, S. J. La Placa, and C. Y. Wong. *J. Appl. Phys.* 55 (1984): 4208.
86. Baglin, J. E. E., H. A. Atwater, D. Gupta, and F. M. d'Heurle. *Thin Solid Films* 93 (1981): 255.
87. Colgan, E. G., and F. M. d'Heurle. *J. Appl. Phys.* 79 (1996): 4087.
88. Lauwers, A., A. Steegen, M. de Potter, R. Lindsay, A. Satta, H. Bender, and K. Maex. *J. Vac. Sci. Technol.* B19 (2001): 2026.
89. Finstad, T. G., J. W. Mayer, and M.-A. Nicolet. *Thin Solid Films* 51 (1978): 391.
90. Tu, K. N. *J. Appl. Phys.* 48 (1977): 3379.
91. Pretorius, R., C. L. Ramiller, S. S. Lau, and M. A. Nicolet. *Appl. Phys. Lett.* 30 (1977): 501.
92. Finstad, T. G. *Phys. Stat. Sol.* A63 (1981): 223.
93. Cheung, N. W., P. J. Grunthaner, F. J. Grunthaner, J. W. Mayer, and B. M. Ullrich. *J. Vac. Sci. Technol.* 18 (1980): 917.
94. van Loenen, E. J., J. F. van der Veen, and F. K. LeGoues. *Surf. Sci.* 157 (1985): 1.
95. Chang, Y.-J., and J. L. Erskine. *J. Vac. Sci. Technol.* A1 (1983): 1193.
96. Foll, H., P. S. Ho, and K. N. Tu. *Philos. Mag.* A45 (1982): 31.
97. Costato, M. *Lettere Al Nuovo Cimento* 32 (1981): 219.
98. Colgan, E. G., M. Maenpaa, M. Finetti, and M.-A. Nicolet. *J. Electron. Mater.* 12 (1983): 413.
99. Schaffer, W. J., R. W. Bene, and R. M. Walser. *J. Vac. Sci. Technol.* 15 (1978): 1325.
100. Teodorescu, V., L. Nistor, H. Bender, A. Steegen, A. Lauwers, K. Maex, and J. Van Landuyt. *J. Appl. Phys.* 90 (2001): 167.
101. Yew, J. Y., H. C. Tseng, L. J. Chen, K. Nakamura, and C. Y. Chang. *Appl. Phys. Lett.* 69 (1996): 3692.
102. Ostling, M., and C. Zaring. *Properties of Metal Silicides*. London: INSPEC, 1995.

103. Lavoie, C., C. Coia, F. M. D'Heurle, C. Detavernier, C. Cabrai Jr., P. Desjardins, and A. J. Kellock. *Defect Diffus. Forum* 237–240 (2005): 825.
104. Bennett, J. M., and L. M. Mattsson. *Introduction to Surface Roughness and Scattering*. Washington: Optical Society of America, 1989.
105. Gas, P., and F. M. d'Heurle. *Appl. Surf. Sci.* 73 (1993): 153.
106. Liew, K. P., R. A. Bernstein, and C. V. Thompson. *J. Mater. Res.* 19 (2004): 676.
107. d'Heurle, F. M., and P. Gas. *J. Mater. Res.* 1 (1986): 205.
108. Zhang, S.-L., and F. M. d'Heurle. *Thin Solid Films* 213 (1992): 34.
109. Gulpen, J., A. A. Kodentsov, and F. J. J. van Loo. *Z. Metallke.* 86 (1995): 530.
110. Madar, R. *Properties of Metal Silicides*. London: INSPEC, 1995.
111. Maex, K. *Mater. Sci. Eng. R: Rep.* R11 (1993): 53.
112. Gas, P., O. Thomas, and F. M. d'Heurle. *Properties of Metal Silicides*. London: INSPEC, 1995.
113. Pawlak, M. A., T. Janssens, A. Lauwers, A. Vantomme, W. Vandervorst, K. Maex, and J. A. Kittl. *Appl. Phys. Lett.* 87 (2005): 181910.
114. Jiang, Y.-L., A. Agarwal, G.-P. Ru, G. Cai, and B.-Z. Li. *Nuclear Instruments and Methods in Physics Research, Section B: Beam Interactions with Materials and Atoms*, 160. Amsterdam: Elsevier, 2005.
115. Zhang, M., J. Knoch, Q. T. Zhao, A. Fox, St.Lenk, and S. Mantl. *Electron. Lett.* 41 (2005): 1085.
116. Tsuchiya, Y., A. Tobioka, O. Nakatsuka, H. Ikeda, A. Sakai, S. Zaima, and Y. Yasuda. *Jpn. J. Appl. Phys. Part 1* 41 (2002): 2450.
117. Morimoto, T., T. Ohguro, S. Momose, T. Iinuma, I. Kunishima, K. Suguro, I. Katakabe, et al. *IEEE Trans. Electron Devices* 42 (1995): 915.
118. Finetti, M., S. Guerri, P. Negrini, A. Scorzoni, and I. Suni. *Thin Solid Films* 130 (1985): 37.
119. Deng, F., R. A. Johnson, P. M. Asbeck, S. S. Lau, W. B. Dubbelday, T. Hsiao, and J. Woo. *J. Appl. Phys.* 81 (1997): 8047.
120. Herner, S. B., K. S. Jones, H.-J. Gossmann, R. T. Tung, J. M. Poate, and H. S. Luftman. *Appl. Phys. Lett.* 68 (1996): 2870.
121. Herner, S. B., H.-J. Gossmann, and R. T. Tung. *Appl. Phys. Lett.* 72 (1998): 2289.
122. Tyagi, A. K., L. Kappius, U. Breuer, H. L. Bay, J. S. Becker, S. Mantl, and H. J. Dietze. *J. Appl. Phys.* 85 (1999): 7639.
123. Fahey, P. M., P. B. Griffin, and J. D. Plummer. *Rev. Mod. Phys.* 61 (1989): 289.
124. Nicolet, M.-A., and S. S. Lau. *Formation and Characterization of Transition-Metal Silicides*. NY: Academic Press, 1983.
125. Gas, P., and F.Md'Heurle. In *Properties of Metal Silicides*. London: INSPEC, 1995.
126. d'Heurle, F., S. Petersson, L. Stolt, and B. Strizker. *J. Appl. Phys.* 53 (1982): 5678.
127. Rivero, C., P. Gergaud, O. Thomas, B. Froment, and H. Jaouen. *Microelectron. Eng.* 76 (2004): 318.
128. Smigelkas, A. D., and E. O. Kirkendall. *Trans. AIME* 171 (1947): 130.
129. Barnes, R. S. *Nature* 166 (1950): 1032.
130. daSilva, L. C. C., and R. F. Mehl. *Trans. AIME* 191 (1951): 155.
131. Kittl, J. A., A. Lauwers, O. Chamirian, M. Van Dal, A. Akheyar, M. De Potter, R. Lindsay, and K. Maex. *Microelectr. Eng.* 70 (2003): 158.
132. Kim, N.-S., H.-S. Cha, N.-K. Sung, H.-H. Ryu, K.-S. Youn, and W.-G. Lee. *J. Vac. Sci. Technol.* A20 (2002): 1171.
133. Chung, J.-H., J.-E. Lee, J.-S. Park, J.-H. Ku, E.-J. Lee, S.-H. Park, S.-T. Kim, J.-L. Nam, and S.-I. Lee. *Advanced Metallization Conference 2000 (AMC 2000)*, 495. San Diego, CA: Materials Research Society, 2000.
134. Kim, Y.-C., J. Kim, J.-H. Choy, J.-C. Park, and H.-M. Choi. *Appl. Phys. Lett.* 75 (1999): 1270.
135. Coulman, B., and E. K. Broadbent. *J. Vac. Sci. Technol.* A5 (1987): 1419.
136. Chau, R., J. Kavalieros, B. Doyle, A. Murthy, N. Paulsen, D. Lionberger, D. Barlage, R. Arghavani, B. Roberds, and M. Doczy. *IEEE International Electron Devices Meeting 2001—Technical Digest*, 29.1.1, 2001.
137. Kolbesen, B. O., and H. Cerva. *Phys. Stat. Sol. B* 222 (2000): 303.
138. Tian, Y., Y.-L. Jiang, Y. Chen, F. Lu, and B.-Z. Li. *Semic. Sci. Technol.* 17 (2002): 83.
139. Rabadanov, M. Kh. , and M. B. Ataev. *Inorg. Mater.* 38 (2002): 120.

140. Wilson, D. F., and O. B. Cavin. *Scr. Metall. Mater.* 26 (1992): 85.
141. Detavernier, C., C. Lavoie, and F. M. d'Heurle. *J. Appl. Phys.* 93 (2003): 2510.
142. Detavernier, C., A. S. Ozcan, J. L. Jordan-Sweet, E. A. Stach, J. Tersoff, F. M. Ross, and C. Lavoie. *Nature* 426 (2003): 641.
143. Tung, R. T. *Mater. Chem. Phys.* 32 (1992): 107.
144. Tung, R. T. *J.Vac. Sci. Technol. A* 7 (1989): 598.
145. Tung, R. T., A. F. J. Levi, J. P. Sullivan, and F. Schrey. *Phys. Rev. Lett.* 66 (1991): 72.
146. Gibson, J. M., J. L. Batstone, R. T. Tung, and F. C. Unterwald. *Phys. Rev. Lett.* 60 (1988): 1158.
147. Seger, J., and S.-L. Zhang. *Thin Solid Films* 429 (2003): 216.
148. Teichert, S., M. Falke, H. Giesler, G. Beddies, and H. J. Hinneberg. *Thin Solid Films* 336 (1998): 222.
149. Feng, Y. Z., and Z. Q. Wu. *J. Mater. Sci. Lett.* 15 (1996): 2000.
150. Sullivan, J. P., R. T. Tung, F. Schrey, and W. R. Graham. *Proccedings of the Semiconductor Heterostructures for Photonic and Electronic Applications Symposium, 30 Nov. to 4 Dec. 1992*, 623. Materials Research Society Symposium Proceedings, 1993.
151. von Kanel, H., T. Graf, J. Henz, M. Ospelt, and P. Wachter. *J. Cryst. Growth* 81 (1987): 470.
152. Byun, J. S., D.-H. Kim, W. S. Kim, and H. J. Kim. *J. Appl. Phys.* 78 (1995): 1725.
153. Alberti, A., F. La Via, C. Spinella, and E. Rimini. *Appl. Phys. Lett.* 75 (1999): 2924.
154. Tung, R. T., A. F. J. Levi, F. Schrey, and M. Anzlowar. "Evaluation of Advanced Semiconductor Materials by Electron Microscopy." In *Proceedings of a NATO Advanced Research Workshop*, 167, 1989.
155. Nygren, S., D. Caffin, M. Ostling, and F. M. d'Heurle. *Appl. Surf. Sci.* 53 (1991): 87.
156. Hong, Q. Z., F. M. d'Heurle, J. M. E. Harper, and S. Q. Hong. *Appl. Phys. Lett.* 62 (1993): 2637.
157. Deduytsche, D., C. Detavernier, R. L. Van Meirhaeghe, and C. Lavoie. *J. Appl. Phys.* 98 (2005): 033526.
158. Zheng, L. R., L. S. Hung, and J. W. Mayer. *Appl. Phys. Lett.* 51 (1987): 2139.
159. Zheng, L. R., L. S. Hung, S. Q. Feng, P. Revesz, J. W. Mayer, and G. Miles. *Appl. Phys. Lett.* 48 (1986): 767.
160. Hong, Q. Z., S. Q. Hong, F. M. D'Heurle, and J. M. E. Harper. *Thin Solid Films* 253 (1994): 479.
161. Mukai, R., S. Ozawa, and H. Yagi. *Thin Solid Films* 270 (1995): 567.
162. Chamirian, O., J. A. Kittl, A. Lauwers, O. Richard, M. van Dal, and K. Maex. *Microelectron. Eng.* 70 (2003): 201.
163. Poon, M. C., M. Chan, W. Q. Zhang, F. Deng, and S. S. Lau. *Microelectron. Rel.* 38 (1998): 1499.
164. Choi, C. J., Y. W. Ok, S. S. Hullavarad, T.-Y. Seong, K.-M. Lee, J.-H. Lee, and Y.-J. Park. *J. Electrochem. Soc.* 149 (2002): G517.
165. Poon, M. C., F. Deng, M. Chan, W. Y. Chan, and S. S. Lau. *Appl. Surf. Sci.* 157 (2000): 29.
166. Tokarev, V. V., A. I. Demchenko, A. I. Ivanov, and V. E. Borisenko. *Appl. Surf. Sci.* 44 (1990): 241.
167. Lee, P. S., D. Mangelinck, K. L. Pey, J. Ding, D. Z. Chi, T. Osipowicz, J. Y. Dai, and A. See. *Microelectron. Eng.* 60 (2002): 171.
168. Lee, P. S., K. L. Pey, D. Mangelinck, J. Ding, D. Z. Chi, J. Y. Dai, and L. Chan. *J. Electrochem. Soc.* 149 (2002): G331.
169. Lee, P. S., K. L. Pey, D. Mangelinck, J. Ding, D. Z. Chi, and L. Chan. *IEEE Electron Devices Lett.* 22 (2001): 568.
170. Mangelinck, D., J. Y. Dai, J. Pan, and S. K. Lahiri. *Appl. Phys. Lett.* 75 (1999): 1736.
171. Liu, J. F., H. B. Chen, and J. Y. Feng. *J. Cryst. Growth* 220 (2000): 488.
172. Liu, J. F., H. B. Chen, J. Y. Feng, and J. Zhu. *Appl. Phys. Lett.* 77 (2000): 2177.
173. Liu, J. F., J. Y. Feng, and J. Zhu. *Appl. Phys. Lett.* 80 (2002): 270.
174. Mangelinck, D., J. Y. Dai, S. K. Lahiri, C. S. Ho, and T. Osipowicz. *Mater. Res. Soc. Symp. Proc.* 564 (1999): 163.
175. Seng, H. L., T. Osipowicz, P. S. Lee, D. Mangelinck, T. C. Sum, and F. Watt. *Nucl. Instr. Meth. Phys. Res. Sect. B* 181 (2001): 399.
176. Lee, P. S., K. L. Pey, D. Mangelinck, J. Ding, A. T. S. Wee, and L. Chan. *IEEE Electron Devices Lett.* 21 (2000): 566.

177. Chen, L. J., S. L. Cheng, S. M. Chang, Y. C. Peng, H. Y. Huang, and L. W. Cheng. *Mater. Res. Soc. Symp. Proc.* 564 (1999): 123.
178. Chao, T.-S., and L.-Y. Lee. *Jpn. J. Appl. Phys. Part 2* 41 (2002): L124.
179. Lee, P. S., D. Mangelinck, K. L. Pey, J. Ding, D. Z. Chi, J. Y. Dai, and A. See. *J. Electron. Mater.* 30 (2001): 1554.
180. Lee, P. S., K. L. Pey, D. Mangelinck, J. Ding, D. Z. Chi, T. Osipowicz, J. Y. Dai, and L. Chan. *J. Electrochem. Soc.* 149 (2002): 505.
181. Chen, W. J., and L. J. Chen. *J. Appl. Phys.* 70 (1991): 2628.
182. Choi, C.-J., Y.-W. Ok, T.-Y. Seong, and H.-D. Lee. *Jpn. J. Appl. Phys. Part 1* 41 (2002): 1969.
183. Wong, A. S. W., D. Z. Chi, M. Loomans, D. Ma, M. Y. Lai, W. C. Tjiu, S. J. Chua, C. W. Lim, and J. E. Greene. *Appl. Phys. Lett.* 81 (2002): 5138.
184. Donthu, S. K., D. Z. Chi, A. S. W. Wong, S. J. Chua, and S. Tripathy. *Mater. Res. Soc. Symp. Proc.* 716 (2002): 465.
185. (a) Juang, M.-H., S.-C. Han, and M.-C. Hu. *Jpn. J. Appl. Phys. Part 1* 37 (1998): 5515; (b) Lavoie, C., C. Cabral, C. Detavernier, and F. M. d'Heurle, J. Jordan-Sweet. "Effects of Additive Elements on the Phase Formation and Morphological Stability of Nickel Monosilicide Films." *Microelectron. Eng.* 83 (2006): 2042–2054.
186. Chi, D. Z., D. Mangelinck, A. S. Zuruzi, A. S. W. Wong, and S. K. Lahiri. *J. Electron. Mater.* 30 (2001): 1483.
187. Welzel, V., M. Leoni, and E. J. Mittemeijer. *Philos. Mag.* 83 (2003): 603.
188. Qin, M., and M. C. P. Vincent. *J. Mater. Sci. Lett.* 19 (2000): 2243.
189. Zielinski, E. M., R. P. Vinci, and J. C. Bravman. *J. Appl. Phys.* 76 (1994): 4516.
190. Thompson, C. V. *Scripta Metall. Mater.* 28 (1993): 167.
191. Thompson, C. V., and R. Carel. *Mater. Sci. Eng.* B32 (1995): 211.
192. Harper, J. M. E., K. P. Rodbell, E. G. Colgan, and R. H. Hammond. *J. Appl. Phys.* 82 (1997): 4319.
193. *International Technology Roadmap for Semiconductors—2004 Update* (http://www.itrs.net/Common/2004Update/2004Update.htm, 2004).
194. Wong, H.-S. P., D. J. Frank, P. M. Solomon, C. H. J. Wann, and J. J. Welser. *Proc. IEEE* 87 (1999): 537.
195. King, T.-J., J. R. Pfiester, J. D. Shott, J. P. McVittie, and K. C. Saraswat. *IEEE International Electron Devices Meeting 1990—Technical Digest*, 253. San Francisco, CA, 1990.
196. Hellberg, P.-E., S.-L. Zhang, and C. S. Petersson. *IEEE Electron Device Lett.* 18 (1997): 456.
197. Yeo, Y.-C., Q. Lu, T.-J. King, C. Hu, T. Kawashima, M. Oishi, S. Mashiro, and J. Sakai. *IEEE International Electron Devices Meeting 2000—Technical Digest*, 753. San Francisco, CA, 2000.
198. Wu, D., A.-C. Lindgren, S. Persson, G. Sjoblom, M. von Haartman, J. Seger, P.-E. Hellstrom, et al. *IEEE Electron Device Lett.* 24 (2003): 171.
199. Mizuno, T., N. Sugiyama, T. Tezuka, and S. Takagi. *IEEE Trans. Electron Devices* 49 (2002): 7.
200. Hoyt, J. L., H. M. Nayfeh, S. Eguchi, I. Aberg, G. Xia, T. Drake, E. A. Fitzgerald, and D. A. Antoniadis. *IEEE International Electron Devices Meeting 2002—Technical Digest*, 23. San Francisco, CA, 2002.
201. Gannavaram, S., N. Pesovic, and C. Ozturk. *IEEE International Electron Devices Meeting 2000—Technical Digest*, 437. San Francisco, CA, 2000.
202. Nishiyama, A., K. Matsuzawa, and S. Takagi. *IEEE Trans. Electron Devices* 48 (2001): 1114.
203. Ozturk, M.C., J. Liu, and H. Mo. *IEEE International Electron Devices Meeting 2003*, 497. Washington, DC, 2003.
204. Ozturk, M.C., J. Liu, H. Mo, and N. Pesovic. *IEEE International Electron Devices Meeting 2002*, 375. San Francisco, CA, 2002.
205. Uchino, T., T. Shiba, K. Ohnishi, A. Miyauchi, M. Nakata, Y. Inoue, and T. Suzuki. *IEEE International Electron Devices Meeting 1997*, 479. Washington, DC, 1997.
206. Huang, H.-J., K.-M. Chen, C.-Y. Chang, L.-P. Chen, G.-W. Huang, and T.-Y. Huang. *IEEE Electron Device Lett.* 21 (2000): 448.
207. Thompson, S. E., M. Armstrong, C. Auth, S. Cea, R. Chau, G. Glass, T. Hoffman, et al. *IEEE Electron Device Lett.* 25 (2004): 191.

208. Lee, S. H., D. S. Shin, H. S. Rhee, T. Ueno, H. Lee, M. H. Park, N.-I. Lee, H.-K. Kang, and K.-P. Suh. *IEEE International Electron Devices Meeting 2004*, 1051. San Francisco, CA, 2004.
209. Zhang, S.-L. *Microelectron. Eng.* 201 (2003): 174.
210. Burnette, J., M. Himmerlich, and R. J. Nemanich. *Silicide Contacts for Si/Ge Devices*, Silicide Technology for Integrated Circuits. London: IEEE, 2004.
211. Zhang, S.-L. "Si-Based CMOS: New Materials, Processes and Equipment, Quebec City, Canada." In *Proceedings of the International Symposium on Advanced Gate Stack and Channel Engineering* , 596. Pennington, U.S.A.: The Electrochemical Society, 2005.
212. Seger, J., S.-L. Zhang, D. Mangelinck, and H. H. Radamson. *Appl. Phys. Lett.* 81 (2002): 1978.
213. Jarmar, T., J. Seger, F. Ericson, D. Mangelinck, U. Smith, and S.-L. Zhang. *J. Appl. Phys.* 92 (2002): 7193.
214. Pey, K. L., W. K. Choi, S. Chattopadhyay, H. B. Zhao, E. A. Fitzgerald, D. A. Antoniadis, and P. S. Lee. *J. Vac. Sci. Technol.* A20 (2002): 1903.
215. Zhao, H. B., K. L. Pey, W. K. Choi, S. Chattopadhyay, E. A. Fitzgerald, D. A. Antoniadis, and P. S. Lee. *J. Appl. Phys.* 92 (2002): 214.
216. Lin, C. Y., W. J. Chen, C. H. Lai, A. Chin, and J. Liu. *IEEE Electron Device Lett.* 23 (2002): 464.
217. Yang, T.-H., G. Luo, E. Y. Chang, T.-Y. Yang, H.-C. Tseng, and C.-Y. Chang. *IEEE Electron Device Lett.* 24 (2003): 544.
218. Wu, W. W., S. L. Cheng, S. W. Lee, and L. J. Chen. *J. Vac. Sci. Technol.* B21 (2003): 2147.
219. Chen, X., Z. Shi, S. K. Banerjee, J. P. Zhou, and L. K. Rabenberg. *J. Electron. Mater.* 32 (2003): 1171.
220. Shi, Z., D. Onsongo, X. Chen, D.-W. Kim, R. E. Nieh, and S. K. Banerjee. *J. Electron. Mater.* 32 (2003): 184.
221. Hallstedt, J., M. Blomqvist, P. O. A. Persson, L. Hultman, and H. H. Radamson. *J. Appl. Phys.* 95 (2004): 2397.
222. Seger, J., T. Jarmar, Z.-B. Zhang, H. H. Radamson, F. Ericson, U. Smith, and S.-L. Zhang. *J. Appl. Phys.* 96 (2004): 1919.
223. Yamamoto, T., A. Sakai, T. Egawa, N. Taoka, O. Nakatsuka, S. Zaima, and Y. Yasuda. *Appl. Surf. Sci.* 224 (2004): 108.
224. Ok, Y.-W., H. S. Kim, Y.-J. Song, K.-H. Shim, and T.-Y. Seong. *Semicond. Sci. Technol.* 19 (2004): 285.
225. Ok, Y.-W., S.-H. Kim, Y.-J. Song, K.-H. Shim, and T.-Y. Seong. *J. Vac. Sci. Technol.* B22 (2004): 1088.
226. Pey, K. L., S. Chattopadhyay, W. K. Choi, Y. Miron, E. A. Fitzgerald, D. A. Antoniadis, and T. Osipowicz. *J. Vac. Sci. Technol.* B22 (2004): 852.
227. Nath, R., and M. Yeadon. *Electrochem. Solid-State Lett.* 7 (2004): 231.
228. Yao, H. B., S. Tripathy, and D. Z. Chi. *Electrochem. Solid-State Lett.* 7 (2004): 323.
229. Wang, G.-W., G.-P. Ru, X.-P. Qu, and B.-Z. Li. *Mater. Lett.* 58 (2004): 2082.
230. Zhao, Q. T., D. Buca, St.Lenk, R. Loo, M. Caymax, and S. Mantl. *Microelectron. Eng.* 76 (2004): 285.
231. Chamirian, O., A. Lauwers, J. A. Kittl, M. van Dal, M. De Potter, R. Lindsay, and K. Maex. *Microelectron. Eng.* 76 (2004): 297.
232. He, Y., X. L. Liu, J. Y. Feng, and Q. L. Wu. *J. Appl. Phys.* 96 (2004): 6928.
233. Seger, J., T. Jarmar, F. Ericson, U. Smith, J. Hallstedt, Z.-B. Zhang, and S.-L. Zhang. *J. Appl. Phys.* 96 (2004): 7179.
234. Jarmar, T., Z.-B. Zhang, J. Seger, F. Ericson, U. Smith, and S.-L. Zhang. *Phys. Rev.* B70 (2004): 235307.
235. Jin, L. J., K. L. Pey, W. K. Choi, E. A. Fitzgerald, D. A. Antoniadis, A. J. Pitera, M. L. Lee, D. Z. Chi, and C. H. Tung. *Thin Solid Films* 462–3 (2004): 151.
236. Jin, L. J., K. L. Pey, W. K. Choi, E. A. Fitzgerald, D. A. Antoniadis, A. J. Pitera, M. L. Lee, and C. H. Tung. *J. Appl. Phys.* 97 (2005): 104917.
237. Takizawa, H., K. Uheda, and T. Endo. *J. Alloys Compd.* 305 (2000): 306.
238. Hsu, S.-L., C.-H. Chien, M.-J. Yang, R.-H. Huang, C.-C. Leu, S.-W. Shen, and T.-H. Yang. *Appl. Phys. Lett.* 86 (2005): 251906.

239. Aldrich, D. B., F. M. D'Heurle, D. E. Sayers, and R. J. Nemanich. *Phys. Rev. B: Condens. Matter* 53 (1996): 16279.
240. Aldrich, D. B., Y. L. Chen, D. E. Sayers, R. J. Nemanich, S. P. Ashburn, and M. C. Ozturk. *J. Appl. Phys.* 77 (1995): 5107.
241. Aldrich, D. B., Y. L. Chen, D. E. Sayers, R. J. Nemanich, S. P. Ashburn, and M. C. Ozturk. *J. Mater. Res.* 10 (1995): 2849.
242. Qi, W.-J., B.-Z. Li, W.-N. Huang, Z.-G. Gu, H.-Q. Lu, X.-J. Zhang, M. Zhang, G.-S. Dong, D. C. Miller, and R. G. Aitken. *J. Appl. Phys.* 77 (1995): 1086.
243. Donaton, R. A., K. Maex, A. Vantomme, G. Langouche, Y. Morciaux, A. St. Amour, and J. C. Sturm. *Appl. Phys. Lett.* 70 (1997): 1266.
244. Tillack, B., P. Zaumseil, G. Morgenstern, D. Krueger, B. Dietrich, and G. Ritter. *J. Cryst. Growth* 157 (1995): 181.
245. Nolan, T. P., R. Sinclair, and R. Beyers. *J. Appl. Phys.* 71 (1992): 720.
246. Thompson, R. D., K. N. Tu, J. Angillelo, S. Delage, and S. S. Iyer. *J. Electrochem. Soc.* 135 (1988): 3161.
247. McVay, G. L, and A. R. DuCharme. *Lattice Defects in Semiconductors*, 91. Freiburg, West Germany: Institute of Physics, 1975.
248. Sunami, H. *J. Electrochem. Soc.* 125 (1978): 892.
249. Sharma, B. L. *Diffusion Defect Data—Solid State Data, Part A (Defect and Diffusion Forum)* 70–71 (1990): 1.
250. Lindgren, A.-C., C. Chen, S.-L. Zhang, M. Ostling, Y. Zhang, and D. Zhu. *J. Appl. Phys.* 91 (2002): 2708.
251. Hong, Q. Z., and J. W. Mayer. *J. Appl. Phys.* 66 (1989): 611.
252. Liou, H. K., X. Wu, U. Gennser, V. P. Kesan, S. S. Iyer, K. N. Tu, and E. S. Yang. *Appl. Phys. Lett.* 60 (1992): 577.
253. Deboer, F. R., R. Boom, W. C. Mattens, A. R. Miedema, and A. K. Niessen. *Cohesion in Metals: Transition Metal Alloys*. Amsterdam, Holland: North Holland, 1988.
254. Detavernier, C., and C. Lavoie. *Appl. Phys. Lett.* 84 (2004): 3549.
255. Qin, M., V. M. C. Poon, and S. C. H. Ho. *J. Electrochem. Soc.* 148 (2001): 271.
256. Lepselter, M. P., and S. M. Sze. *IEEE Proc.* 56 (1968): 1400.
257. Kedzierski, J., P. Xuan, E. H. Anderson, J. Bokor, T.-J. King, and C. Hu. *IEEE International Electron Devices Meeting 2000*, 57. San Francisco, CA, 2000.
258. Xu, Z. *Properties of Metal Silicides*. London: INSPEC, 1995.
259. Derrien, J. *Properties of Metal Silicides*. London: INSPEC, 1995.
260. Tove, P. A., K. Bohlin, F. Masszi, H. Norde, J. Nylander, J. Tiren, and U. Magnusson. *IEEE Electron Device Lett.* 9 (1988): 47.
261. Hillenius, S. J., R. Liu, G. E. Georgiou, R. L. Field, D. S. Williams, A. Kornblit, D. M. Boulin, R. L. Johnston, and W. T. Lynch. *IEEE International Electron Devices Meeting 1986*, 252. Los Angeles, CA, 1986.
262. Hu, G. J., and R. H. Bruce. *IEEE Trans. Electron Devices* ED-32 (1985): 584.
263. Wilk, G. D., R. M. Wallace, and J. M. Anthony. *J. Appl. Phys.* 89 (2001): 5243.
264. Rios, R., and N. D. Arora. *IEEE International Electron Devices Meeting 1994*, 613. San Francisco, CA, 1994.
265. Tung, R. T. *Mater. Sci. Eng.: R: Rep.* 35 (2001): 138.
266. Wakabayashi, H., Y. Saito, K. Takeuchi, T. Mogami, and T. Kunio. *IEEE International Electron Devices Meeting 1999*, 253. Washington, DC, 1999.
267. Lin, R., Q. Lu, P. Ranade, T.-J. King, and C. Hu. *IEEE Electron Device Lett.* 23 (2002): 49.
268. Polishchuk, I., P. Ranade, T.-J. King, and C. Hu. *IEEE Electron Device Lett.* 23 (2002): 200.
269. Kedzierski, J., E. Nowak, T. Kanarsky, Y. Zhang, D. Boyd, R. Carruthers, C. Cabrai et al. *IEEE International Electron Devices Meeting 2002*, 247. San Francisco, CA, 2002.
270. Maszara, W. P. *J. Electrochem. Soc.* 152 (2005): 550.
271. Pezzi, R. P., M. Copel, C. Cabral, and I. J. R. Baumvol. *Appl. Phys. Lett.* 87 (2005): 162902.
272. Copel, M., R. P. Pezzi, and C. Cabral Jr.. *Appl. Phys. Lett.* 86 (2005): 251904.

273. Wen, H. C., J. Liu, J. H. Sim, J. P. Lu, and D. L. Kwong. *Electrochem. Solid-State Lett.* 8 (2005): 119.
274. Xuan, P., and J. Bokor. *IEEE Electron Device Lett.* 24 (2003): 634.
275. Liu, J., H. C. Wen, J. P. Lu, and D.-L. Kwong. *IEEE Electron Device Lett.* 26 (2005): 228.
276. Lee, R. T. P., S. L. Liew, W. D. Wang, E. K. C. Chua, S. Y. Chow, M. Y. Lai, and D. Z. Chi. *Electrochem. Solid-State Lett.* 8 (2005): 156.
277. Yeo, Y.-C., P. Ranade, T.-J. King, and C. Hu. *IEEE Electron Device Lett.* 23 (2002): 242.
278. Yeo, Y.-C., T.-J. King, and C. Hu. *J. Appl. Phys.* 92 (2002): 7266.
279. Heine, V. *Phys. Rev.* 138 (1965): A1689–A96.
280. Yu, H. Y., C. Ren, Y.-C. Yeo, J. F. Kang, X. P. Wang, H. H. H. Ma, M.-F. Li, D. S. H. Chan, and D.-L. Kwong. *IEEE Electron Device Lett.* 25 (2004): 337.
281. Nicolet, M.-A., and S. S. Lau. *VLSI Electronics—Microstructure Sciences*. New York: Academic Press, 1983.
282. Tu, K. N., R. D. Thompson, and B. Y. Tsaur. *Appl. Phys. Lett.* 38 (1981): 626.
283. Norde, H., J. de Sousa Pires, F. d'Heurle, F. Pesavento, S. Petersson, and P. A. Tove. *Appl. Phys. Lett.* 38 (1981): 865.
284. Zhu, S., J. Chen, M.-F. Li, S. J. Lee, J. Singh, C. X. Zhu, A. Du, C. H. Tung, A. Chin, and D. L. Kwong. *IEEE Electron Device Lett.* 25 (2004): 565.
285. Tucker, J. R., C. Wang, and P. S. Carney. *Appl. Phys. Lett.* 65 (1994): 618.
286. Zhu, S., H. Y. Yu, S. J. Whang, J. H. Chen, C. Shen, C. Zhu, S. J. Lee., et al. *IEEE Electron Device Lett.* 25 (2004): 268.
287. Ikeda, K., Y. Yamashita, A. Endoh, T. Fukano, K. Hikosaka, and T. Mimura. *IEEE Electron Device Lett.* 23 (2002): 670.
288. Larrieu, G., and E. Dubois. *IEEE Electron Device Lett.* 25 (2004): 801.
289. Lin, H.-C., M.-F. Wang, F.-J. Hou, H.-N. Lin, C.-Y. Lu, J.-T. Liu, and T.-Y. Huang. *IEEE Electron Device Lett.* 24 (2003): 102.
290. Jang, M., Y. Kim, J. Shin, S. Lee, and K. Park. *Appl. Phys. Lett.* 84 (2004): 741.
291. Deng, F., K. Ring, Z. F. Guan, S. S. Lau, W. B. Dubbelday, N. Wang, and K. K. Fung. *J. Appl. Phys.* 81 (1997): 8040.
292. Deng, F., R. A. Johnson, P. M. Asbeck, S. S. Lau, W. B. Dubbelday, T. Hsiao, and J. Woo. *J. Appl. Phys.* 81 (1997): 8047.
293. Seger, J., P.-E. Hellstrom, J. Lu, B. G. Malm, M. von Haartman, M. Ostling, and S.-L. Zhang. *Appl. Phys. Lett.* 86 (2005): 253507.
294. Wang, J., E. Polizzi, and M. Lundstrom. *IEEE International Electron Devices Meeting 2003*, 695. Washington, DC, 2003.
295. King, T.-J. *Electrochem. Soc. Interface* 14 (2005): 38.
296. Kedzierski, J., M. Ieong, E. Nowak, T. S. Kanarsky, Y. Zhang, R. Roy, D. Boyd, D. Fried, and H.-S. P. Wong. *IEEE Trans. Electron Devices* 50 (2003): 952.
297. Rhoderick, E. H., and R. H. Williams. *Metal-Semiconductor Contacts*. Oxford, England: Clarendon Press, 1988.

11

Rapid Thermal Processing

11.1 Introduction **11**-1

11.2 RTP System Hardware and Control Technology **11**-1
RTP System Configurations • Thermal Radiation Physics • Calculation of Thermal Radiative Properties • Optical Properties of Materials Present in Wafers • Thermal Radiative Properties of Semiconductor Wafers • Thermal Response of Wafers • Temperature Measurement and Control in RTP • Process Uniformity Control in RTP

11.3 Semiconductor Processing Using RTP **11**-74
RTP Applications in Dielectric Formation and Processing • Applications of RTP in Ion Implantation Damage Annealing and Dopant Activation • Applications of RTP in Forming Contacts and Interconnect Structures • Emerging Applications of RTP: Strain Engineering, Metal Gates, and Multi-Gate CMOS • Applications of RTP beyond CMOS

11.4 Conclusion **11**-101

Acknowledgments **11**-101

References **11**-101

P. J. Timans
Mattson Technology

11.1 Introduction

Rapid thermal processing (RTP) is a key technology in the fabrication of advanced integrated circuits, with a wide range of applications, including metal silicide and nitride formation, ion implantation damage annealing and activation, dielectric formation and annealing, and the reflow of deposited oxides [1,2]. Table 11.1a and Table 11.1b summarize some of the current and future process applications for RTP. The technology of RTP is radically different to that of conventional furnace systems. Typical RTP systems use radiant energy sources, often tungsten–halogen lamps, to heat a wafer to a high temperature for a period lasting less than a minute. Shrinking device dimensions and increasing wafer diameters are expected to make the use of RTP even more widespread, as a result of its low thermal budget, fast cycle time, and compatibility with single-wafer processing. New applications, including gate dielectric formation and rapid thermal chemical vapor deposition (RTCVD), are also emerging.

11.2 RTP System Hardware and Control Technology

The main objective of RTP system design and operation is tight temperature control. As semiconductor devices decrease in size, it becomes increasingly important to minimize temperature variations that can

TABLE 11.1a Applications for Rapid Thermal Processing (RTP) in Silicon Device Technology

Node	65 nm	45 nm	32 nm	22 nm
Volume Production	2005	2007	2010	2014
Leading Technology	Bulk Si	Bulk Si	FD–SOI	FD–SOI/Multigate
Dielectrics	SiON form/anneal	SiON form/anneal	SiON form/anneal	
	Surface nitridation	Surface nitridation	Surface nitridation	Surface nitridation
	Post-nitridation SiON anneal	Post-nitridation SiON anneal	Post-nitridation SiON anneal	Post-nitridation High-k anneal
	Sidewall oxide	Sidewall oxide	Sidewall oxide	Sidewall oxide
	Selective oxide (W-gate DRAM)	Selective oxide (W-gate DRAM)	Selective oxide (W-gate DRAM)	
	Sacrificial oxide	Sacrificial oxide	Sacrificial oxide	Sacrificial oxide
	Pad oxide	Pad oxide	Pad oxide	Pad oxide
	Shallow-trench isolation (STI) liner oxide	STI liner oxide	STI liner oxide	STI liner oxide
	Tunnel oxide	Tunnel oxide/nitride	Tunnel oxide/nitride	Tunnel dielectric
	Interpoly oxide	Interpoly oxide	Interpoly dielectric	Interpoly dielectric
	Deposited oxide densification	Deposited oxide densification	Deposited oxide densification	Deposited oxide densification
		High-*k* interface layer engineering	High-*k* interface layer engineering	High-*k* interface layer engineering
		High-*k* post-deposition anneal (PDA)	High-*k* PDA	High-*k* PDA
				Multi-gate corner shape engineering
				Multi-gate channel surface smoothing
Doping processes	Source/Drain extension anneal	Source/Drain extension anneal	Source/Drain extension anneal	Source/Drain anneal
	"Deep" s/d anneal	"Deep" s/d anneal	Raised s/d anneal	Raised s/d anneal
	Well and Channel implant anneal	Well and Channel implant anneal		
Gate electrode processes	Poly-Si/SiGe gate activation	Poly-Si/SiGe gate activation	Poly-SiGe gate activation	
		Metal gate work-function tuning	Metal gate work-function tuning	Metal gate work-function tuning
		Full silicidation of polysilicon (FUSI)	Full silicidation of polysilicon (FUSI)	Full silicidation of polysilicon (FUSI)
Contacts	$CoSi_2$ form/anneal			
	NiSi/Ni SiGe formation and anneal	NiSi/Ni(Pt) SiGe formation and anneal	NiSi/Ni(Pt) SiGe/NiGe formation and anneal	NiSi/Ni(Pt) SiGe/NiGe formation and anneal
			Dual silicide formation and anneal	Dual silicide formation and anneal
				Schottky S/D form
Interconnect	Barrier layer anneal	Barrier layer anneal	Barrier layer anneal	Barrier layer anneal
		Cu anneal	Cu anneal	Cu anneal
		Low-*k* curing	Low-*k* curing	Low-*k* curing

Source: Reprinted From MacKnight, R. B., Timans, P. J., Tay, S.-P., and Nenyei, Z., in *12th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2004*, 3. With permission (©2004 IEEE.)

TABLE 11.1b Applications for RTP beyond Conventional Silicon Technology

Volume production	2005	2007	2010	2013
Wafer manufacturing	Thermal donor annihilation	Thermal donor annihilation	Thermal donor annihilation	
	Magic Denuded Zone® [127]	Magic Denuded Zone® [127]	Magic Denuded Zone® [127]	
	COP Anneal	COP Anneal	COP Anneal	
	SOI surface smoothing	SOI surface smoothing	SOI surface smoothing	SOI surface smoothing
	Strained Si and SOI substrates	Strained Si and SOI substrates	Strained Si and SOI substrates	Strained Si and SOI substrates
Compound semiconductors	Contact annealing	Contact annealing	Contact annealing	Contact annealing
	Dopant activation: p-GaN	Dopant activation: p-GaN	Dopant activation: p-GaN	Dopant activation: p-GaN
	Implant anneal: GaAs	Implant anneal: SiC related	Implant anneal: SiC related	Implant anneal: SiC related
		Quantum well intermixing	Quantum well intermixing	Quantum well intermixing
		Selective oxidation for VCSEL	Selective oxidation for VCSEL	Selective oxidation for VCSEL
Si-based optoelectronics		Waveguide engineering	Waveguide engineering	Waveguide engineering
			Nanoparticle formation	Nanoparticle formation
			CMOS integrated optoelectronics (incl on-chip comms)	CMOS integrated optoelectronics (incl on-chip comms)
Solar cell processes	Doping, oxidation, contacts	Doping, oxidation, contacts	Doping, oxidation, contacts	Doping, oxidation, contacts
Flat-panel displays	Crystallization, doping, oxidation, contacts	Crystallization, doping, oxidation, contacts	Crystallization, doping, oxidation, contacts	Crystallization, doping, oxidation, contacts
Data storage	Magnetic film annealing	Magnetic film annealing	Magnetic film annealing	Magnetic film annealing
	FRAM anneal	FRAM anneal	FRAM anneal	FRAM anneal
		Ovonic memory processes	Ovonic memory processes	Ovonic memory processes
		MRAM processes	MRAM processes	MRAM processes
				Single-electron memory fabrication
Passive components	High-value capacitors	High-value capacitors	High-value capacitors	High-value capacitors
MEMS fabrication	Stress relief annealing	Stress relief annealing	Stress relief annealing	Stress relief annealing
				Doping, oxidation, contacts

Source: Reprinted From MacKnight, R. B., Timans, P. J., Tay, S.-P., and Nenyei, Z., in *12th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2004*, 3. With permission (©2004 IEEE.)

cause fluctuations in the diffusion of atoms or in film thicknesses. For fabrication of advanced devices, process temperatures typically need to be controlled to within ± 2°C [3].

11.2.1 RTP System Configurations

Figure 11.1 shows some typical RTP system configurations that have been developed commercially. Figure 11.1a shows the most common configuration, where a wafer in a quartz envelope is irradiated from both above and below by banks of linear tungsten–halogen lamps [4,5]. The quartz tube is usually

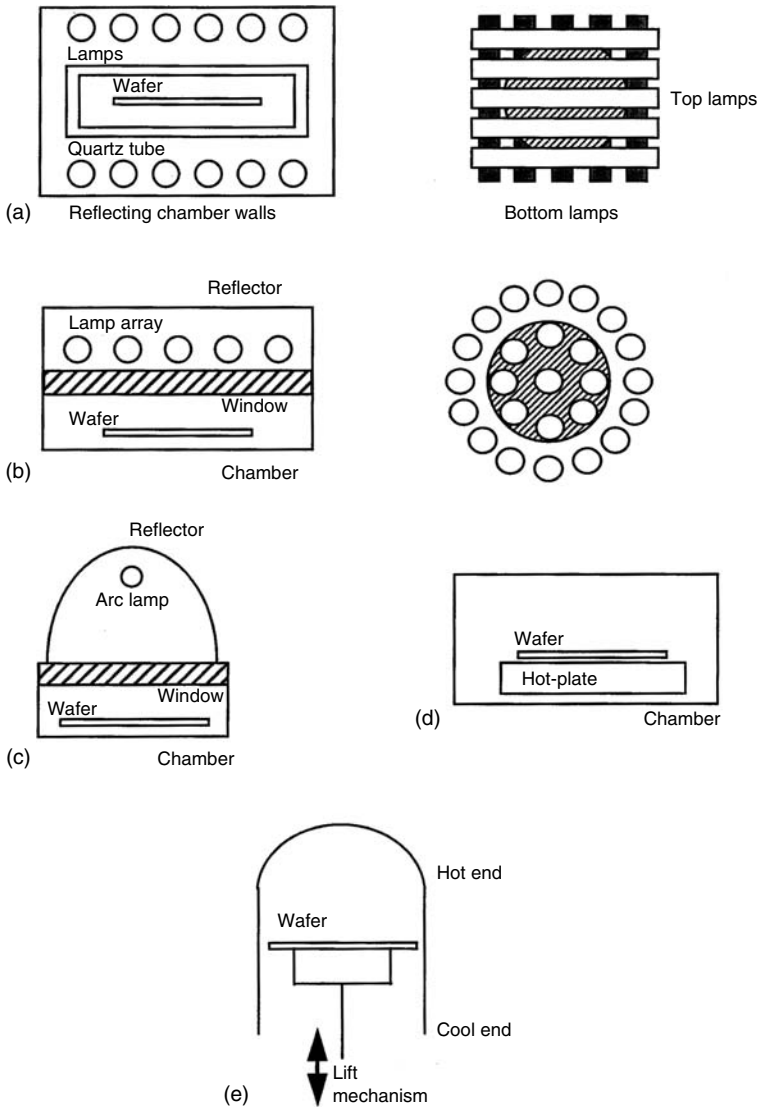


FIGURE 11.1 Typical rapid thermal processing (RTP) tool configurations. (a) Linear lamps, double-sided heating; (b) axisymmetric lamps, single-sided heating; (c) arc lamp; (d) hot-plate; and (e) hot-wall RTP.

cooled by jets of air, and hence the system could be described as a "warm-wall" system. A second type of system, shown in Figure 11.1b, irradiates the wafer from only one side, using an axisymmetric lamp array, while the wafer faces a reflecting surface on the opposite side [6–8]. The wafer is isolated from the lamps by a cooled quartz window. In this type of system, the use of the axisymmetric lamp array is often combined with wafer rotation to improve wafer temperature uniformity. A third kind of lamp-heated system, shown in Figure 11.1c, uses a single arc lamp as the primary power source [4,9]. Figure 11.1d shows a non-lamp configuration in which the wafer is loaded onto pins above a hot-plate which is kept at a constant temperature above the desired process temperature [10,11]. Another non-lamp system is shown in Figure 11.1e, where the wafer is heated by elevating it into a hot-wall chamber that is kept at a temperature above the desired process temperature, which is determined by the position to which the

wafer is raised in the furnace [12]. Many other types of RTP system have been developed since interest in the subject started in the late 1970s, and RTCVD systems are also available commercially.

11.2.2 Thermal Radiation Physics

Many important differences between RTP and conventional furnace processing arise from the fact that in RTP the wafer is generally not in thermal equilibrium with its environment. In a furnace, the wafers are in an isothermal environment, where the furnace walls are at the same temperature as the wafers, whereas in an RTP system the tungsten–halogen lamps are much hotter than the wafer, and the chamber walls are usually much cooler than the wafer. It is these temperature differences that permit rapid heating and cooling of the wafer, but they also have important consequences for the methods of temperature measurement and control that lie at the heart of RTP technology [13]. Because of the high temperatures of both the wafer and the lamps, the physics of RTP is dominated by radiation heat transfer, and the optical properties of the wafer and the chamber play an important part in the behavior [14]. The sections that follow will describe the essential features of the physics of thermal radiation and optics that can be used to understand the most important aspects of temperature control in RTP.

11.2.2.1 Basic Laws

The spectrum of radiation emitted from a blackbody is described by the Planck's radiation function,

$$W_{bb}(\lambda,T) = \frac{c_1}{\lambda^5(\exp(c_2/\lambda T)-1)}, \tag{11.1}$$

where $W_{bb}(\lambda,T)$, the spectral radiant exitance, describes the power per unit area and wavelength radiated into the forward hemisphere from a blackbody at the absolute temperature T in K, at the wavelength λ in μm [15]. The equation gives $W_{bb}(\lambda,T)$ in units of $\mathrm{Wm}^{-2}\mu\mathrm{m}^{-1}$, and c_1 and c_2 are constants, with the values 3.7418×10^8 $\mathrm{W}\mu\mathrm{m}^4\ \mathrm{m}^{-2}$ and 1.4388×10^4 μm K, respectively. Many of the characteristics of RTP systems are consequences of the behavior described by this equation. For example, one is interested in the spectra of radiation from the lamps, the wafer, and the environment. Typically, the tungsten filament in a lamp is at temperature between 1500 and 2500°C, the wafer is between 300 and 1200°C, and the chamber is between 20 and 500°C. Significant variations can arise between the various types of RTP system. For example, in a system using an arc lamp, the effective source temperature is ∼6000°C. Figure 11.2 shows

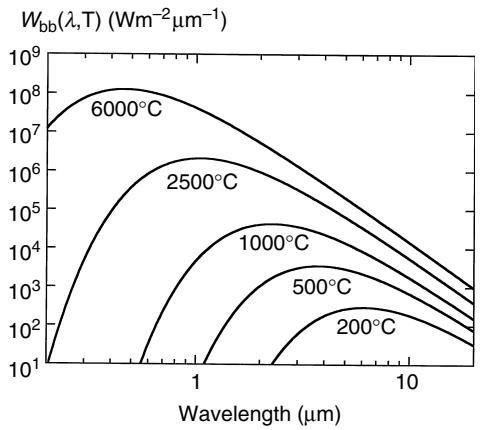


FIGURE 11.2 Blackbody spectra for various source temperatures. $W_{bb}(\lambda,T)$ is the spectral radiant exitance, which describes the power per unit area and wavelength radiated into the forward hemisphere.

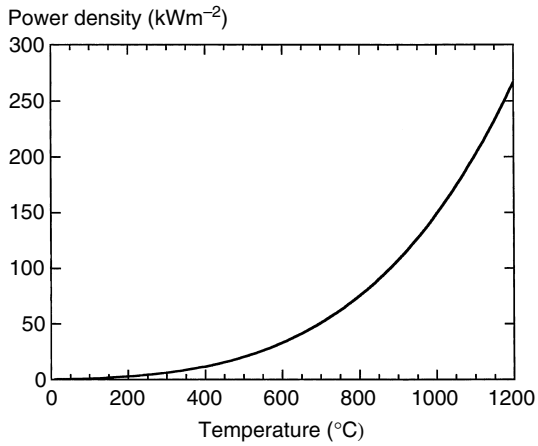


FIGURE 11.3 The Stefan–Boltzmann radiation law describing the total power radiated per unit area of a blackbody as a function of its temperature.

blackbody spectra at various temperatures. The wavelength of the peak of a given spectrum, λ_{max}, can be predicted from Wien's displacement law,

$$\lambda_{max} = \frac{2898}{T}, \tag{11.2}$$

where λ_{max} is in μm and T is in K.

By integrating Equation 11.1 with respect to wavelength, one obtains the Stefan–Boltzmann radiation law,

$$W_{tot,bb}(T) = \sigma T^4, \tag{11.3}$$

where $W_{tot,bb}(T)$ is the total power radiated per unit area, σ is the Stefan–Boltzmann constant (5.67×10^{-8} Wm^{-2} K^{-4}) and T is in K. Figure 11.3 shows the curve corresponding to Equation 11.3.

11.2.2.2 Thermal Radiative Properties

An appreciation of the thermal radiative properties of semiconductor wafers can help one to understand many important phenomena in RTP [14]. The radiation laws given above apply to a blackbody, which is a perfect emitter. Real materials emit less radiation than a blackbody would at any given temperature. The emissivity of an object is the ratio of the radiation it emits to that emitted by a blackbody at the same temperature. The thermal radiation emitted from an object depends on the direction in which the radiation is viewed, its state of polarization and the wavelength, as well as the optical properties of the surface. Many different definitions of emissivity can be constructed with directional and spectral qualifiers, which specify the conditions under which they apply [16].

11.2.2.2.1 Spectral Emissivity and Kirchhoff's Law

The most fundamental thermal radiative property is the spectral emissivity, $\varepsilon(\lambda,\theta,\phi,T)$, which is the ratio of the radiation emitted by a wafer with temperature T, at a given wavelength, λ, angle of incidence, θ, and plane of polarization, ϕ, to that emitted from a blackbody under the same conditions. It is also depends on the azimuthal angle if the surface does not have azimuthal symmetry. This dependence will be ignored here, although it may affect the nature of the radiation emitted from patterned surfaces. The spectral emissivity of an object is identical to its absorptivity, $a(\lambda,\theta,\phi,T)$, for the same conditions. This relationship is known as Kirchhoff's law and can be written as

$$a(\lambda,\theta,\phi,T) = \varepsilon(\lambda,\theta,\phi,T). \tag{11.4}$$

11.2.2.2.2 Integrated Optical Properties

Thermal modeling of RTP systems requires calculation of power coupling and heat loss from wafers. Since radiation heat transfer occurs over a range of wavelengths and depends on the system geometry, the power absorption and thermal emission from the wafer have to be described by quantities that average over the relevant ranges of wavelength and angle of incidence. The relevant spectral and angular ranges depend on the geometry of the system and the nature of the heat source.

For example, in free space, the total radiation heat loss from a surface is governed by the total hemispherical emissivity, $\varepsilon_{\text{tot}}(T)$, which is related to the spectral emissivity by the equation,

$$\varepsilon_{\text{tot}}(T) = \frac{2\int_{\theta=0}^{\pi/2}\int_{\lambda=0}^{\infty}\varepsilon_u(\lambda,\theta,T)W_{\text{bb}}(\lambda,T)\sin\theta\cos\theta\,\mathrm{d}\theta\,\mathrm{d}\lambda}{\sigma T^4}, \tag{11.5}$$

where the term in the denominator is the total power radiated by a blackbody, as given by the Stefan–Boltzmann law. $W_{\text{bb}}(\lambda,T)$ is the spectral radiant exitance and $\varepsilon_{\text{u}}(\lambda,\theta,T)$ is the spectral emissivity for unpolarized radiation. It is important to realize that Kirchhoff's law does not, in general, mean that integrated emissivities can be equated with similar absorptivities. For example, the total hemispherical absorptivity of the wafer with respect to a lamp source, $a_{\text{tot}}(T,T_{\text{L}})$, could be calculated from

$$a_{\text{tot}}(T,T_{\text{L}}) = \frac{\int_{\theta=0}^{\pi/2}\int_{\lambda=0}^{\infty}\varepsilon_u(\lambda,\theta,T)I_{\text{Lamp}}(\lambda,\theta,T_{\text{L}})\sin\theta\cos\theta\,\mathrm{d}\theta\,\mathrm{d}\lambda}{\int_{\theta=0}^{\pi/2}\int_{\lambda=0}^{\infty}I_{\text{Lamp}}(\lambda,\theta,T_{\text{L}})\sin\theta\cos\theta\,\mathrm{d}\theta\,\mathrm{d}\lambda}, \tag{11.6}$$

where $I_{\text{Lamp}}(\lambda,\theta,T_{\text{L}})$ describes the lamp radiation's spectral and directional qualities, and T_{L} is the lamp filament temperature. The equation reflects the fact that the spectral and directional qualities of the lamp radiation may be quite different from those of a blackbody at the wafer temperature, and $\varepsilon_{\text{tot}}(T)$ does not necessarily equal $a_{\text{tot}}(T,T_{\text{L}})$. For the purposes of modeling radiation heat transfer in real RTP systems, other spectral and angular ranges may be more appropriate than those in Equation 11.5 and Equation 11.6.

11.2.3 Calculation of Thermal Radiative Properties

The fundamental optical properties of the materials in a wafer can be used to calculate the spectral emissivity and other properties of interest described above. The optical response of a material can be described by a number of different properties, which are interrelated. The response of a material to electromagnetic radiation is described by the dielectric "constant," ε_{r}, which is actually a function of frequency. The absorption of radiation is included by making ε_{r} a complex quantity. ε_{r} is not usually measured directly, and a number of other properties are used to describe the optical response. The complex refractive index, n_{c}, is defined by the relation

$$n_{\text{c}} = \sqrt{\varepsilon_{\text{r}}} = n - jk, \tag{11.7}$$

where n is usually called the refractive index, and k is the extinction coefficient. Power loss in a material is related to the absorption coefficient, α, which is defined as

$$\alpha = 4\pi k/\lambda, \tag{11.8}$$

where λ is the wavelength. α determines the penetration depth of radiation in a given medium, since the intensity of the radiation decreases according to $\exp(-\alpha z)$, where z is the depth beneath the surface of the medium. α is usually quoted in cm^{-1}. The normal incidence reflectivity, R, of a material is also often measured in experiments. It is given by

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2}. \tag{11.9}$$

11.2.3.1 Calculation of Spectral Emissivity

It is often useful to be able to calculate the spectral emissivity of a coated slab, because this approach can be used to predict the thermal radiative properties of wafers or other objects. The necessary theory has been described in detail elsewhere, and only the main aspects will be described here [17,18]. Figure 11.4 shows a slab with coatings on both its top and bottom surfaces, which has a flux of radiation incident on it. Kirchhoff's law tells us that the absorptivity, which is the fraction of the incident power that is neither reflected nor transmitted, equals the emissivity for radiation emitted at the same wavelength, angle of incidence, and polarization. The fraction of the incident radiation reflected from the top surface is R^*, the apparent reflectivity of the sample, and the fraction transmitted is T^*, the apparent transmissivity. These quantities include the effects of multiple reflections within the slab. Separate calculations of the spectral emissivity are needed for radiation polarized parallel to the plane of incidence (p-polarization or TM wave) and for that polarized perpendicular to this plane (s-polarization or TE wave). The spectral emissivities, for the p- and s-polarizations, are given by

$$\varepsilon_{p(s)} = 1 - R^*_{p(s)} - T^*_{p(s)}, \tag{11.10}$$

where the spectral, angular, and temperature qualifiers are omitted, in order to simplify the notation. Summation of the contributions from the multiply-reflected rays shown in Figure 11.4 gives expressions for R^* and T^*,

$$R^*_{p(s)} = R_{tv} + \frac{a^2 T_t^2 R_{bs}}{1 - a^2 R_{ts} R_{bs}} \tag{11.11}$$

and

$$T^*_{p(s)} = \frac{a T_t T_b}{1 - a^2 R_{ts} R_{bs}}, \tag{11.12}$$

where the $p(s)$ subscript on R^* and T^* indicates that all the quantities in these equations have to take values appropriate for either the p- or s-polarizations. T_t and T_b are the transmissivities of the top and bottom stacks of films; R_{tv} is the reflectivity of the top stack for radiation incident from the outside (vacuum) side and R_{ts} is that for radiation incident from the inside (substrate) side; R_{bs} is the corresponding reflectivity for the bottom stack. a describes the attenuation of intensity experienced by a ray passing through the substrate, which is given by

$$a = \exp(-\alpha d/\cos\theta), \tag{11.13}$$

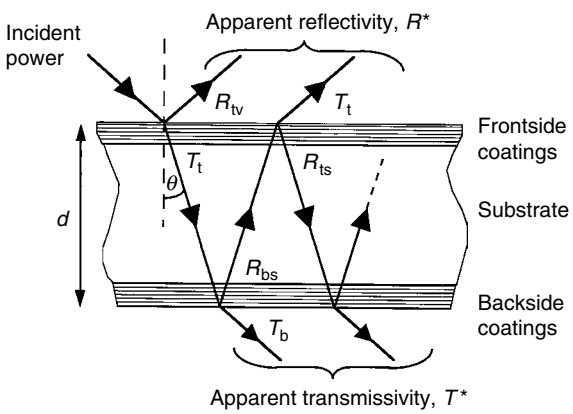


FIGURE 11.4 A general model for a coated substrate of thickness, d, with coatings on both of its surfaces. The arrows show possible paths for radiation incident on the slab and symbols refer to the reflectivities and transmissivities of the top and bottom surfaces.

where d is the thickness of the substrate, α is the absorption coefficient, and θ is the internal angle of propagation. The emissivities can be obtained from Equation 11.10, giving the result

$$\varepsilon_{p(s)} = 1 - R_{tv} - \frac{aT_t(T_b + aT_tR_{bs})}{1 - a^2R_{ts}R_{bs}}. \tag{11.14}$$

The emissivity for unpolarized light is

$$\varepsilon = \frac{\varepsilon_p + \varepsilon_s}{2}. \tag{11.15}$$

If the sample is opaque, $a = 0$, and the spectral emissivity is $(1 - R_{tv})$. The various reflectivities and transmissivities in the right-hand side of Equation 11.14 can be calculated using the theory of thin-film coatings. The relationships between the tangential electric and magnetic fields at the two interfaces of the mth film in a multiple stack can be summarized in matrix notation, as

$$\begin{pmatrix} E_I \\ H_I \end{pmatrix} = \begin{pmatrix} \cos\delta_m & j\sin\delta_m/\gamma_{m(p,s)} \\ j\gamma_{m(p,s)}\sin\delta_m & \cos\delta_m \end{pmatrix} \begin{pmatrix} E_{II} \\ H_{II} \end{pmatrix}, \tag{11.16}$$

where E_I and H_I are the tangential electric and magnetic fields at the first surface of the film, and E_{II} and H_{II} are the analogous quantities at the second surface. δ_m is the phase change for the wave traversing the film at an angle of θ_m to the normal. If the film thickness is d_m and its complex refractive index is n_m, then

$$\delta_m = 2\pi n_m d_m \cos\theta_m/\lambda, \tag{11.17}$$

where λ is the free-space wavelength of the radiation. The expression for $\gamma_{m(p,s)}$ depends on the plane of polarization of wave. For p-polarized radiation,

$$\gamma_{m(p)} = \frac{n_m}{Z_0\cos\theta_m}, \tag{11.18}$$

and for s-polarized radiation,

$$\gamma_{m(s)} = \frac{n_m\cos\theta_m}{Z_0}, \tag{11.19}$$

where Z_0 is the impedance of free space, $\sim 377\ \Omega$. The matrix in Equation 11.16 is called the characteristic matrix of the mth layer. The relationship between the fields at the surface of the uppermost film, and those at the interface between the bottom film and the substrate can be determined by multiplying these matrices together. If there are NF films, then

$$\begin{pmatrix} E_I \\ H_I \end{pmatrix} = M_1M_2\cdots M_{NF} \begin{pmatrix} E_{NF+1} \\ H_{NF+1} \end{pmatrix}, \tag{11.20}$$

where M are the characteristic matrices of the films in the stack. The product of these matrices gives a characteristic matrix for the whole stack, $M_{stack(p,s)}$,

$$M_{stack(p,s)} = \begin{pmatrix} m_{11(p,s)} & m_{12(p,s)} \\ m_{21(p,s)} & m_{22(p,s)} \end{pmatrix}, \tag{11.21}$$

and the elements of this matrix can be used to obtain the reflectivity and transmissivity for the stack. The reflectivity is given by

$$R_{(p,s)} = \left| \frac{(m_{11(p,s)} + \gamma_{\mathrm{sub}(p,s)} m_{12(p,s)})\gamma_{0(p,s)} - m_{21(p,s)} - \gamma_{\mathrm{sub}(p,s)} m_{22(p,s)}}{(m_{11(p,s)} + \gamma_{\mathrm{sub}(p,s)} m_{12(p,s)})\gamma_{0(p,s)} + m_{21(p,s)} + \gamma_{\mathrm{sub}(p,s)} m_{22(p,s)}} \right|^2, \tag{11.22}$$

and the transmissivity is given by

$$T_{(p,s)} = \frac{4\ \mathrm{Re}(\gamma_{0(p,s)})\mathrm{Re}(\gamma_{\mathrm{sub}(p,s)})}{|(m_{11(p,s)} + \gamma_{\mathrm{sub}(p,s)} m_{12(p,s)})\gamma_{0(p,s)} + m_{21(p,s)} + \gamma_{\mathrm{sub}(p,s)} m_{22(p,s)}|^2}. \tag{11.23}$$

In these expressions, $\gamma_{0(p,s)}$ and $\gamma_{\mathrm{sub}(p,s)}$ are given by the Equation 11.18 and Equation 11.19 above, putting the complex refractive index and the angle of propagation equal to the values appropriate for the incident medium and the substrate, respectively. The theory presented here is only strictly valid if the incident medium is non-absorbing. However, for most practical applications, the error introduced by this approach is usually negligible.

11.2.3.2 Effects of Surface Roughness and Patterns

Real wafers may not match the simple model of Figure 11.4, because their surfaces may not be perfectly smooth. The front surface may be textured by patterns etched in coatings, and the back surface may be rough. Under these circumstances, the approach outlined above may not predict the optical properties correctly. If the patterns on the wafer are large compared with the wavelength of the radiation, then geometrical optics can be used, but if their length scale is similar to the wavelength then a full solution of the electromagnetic field problem is required, which requires much more complex methods than those described above. The radiative properties may even lose the azimuthal symmetry with respect to the wafer if there is a tendency toward "grating" behavior. Experimental studies of patterned wafers suggest that simple film models are still useful in predicting the trends of behavior, but that the details differ [19].

Surface roughness also has a complex effect on thermal radiative properties, which depends on the type of roughness and the optical properties of the material. The amplitude, shape, and length scale of the surface texture influence the optical response. Various theories have been developed to deal with simple cases, in particular for metals, where radiation interacts mainly with the surface. For insulators, the behavior is markedly different because the absorption coefficient is usually much lower than it is for metals and thermal radiation is emitted from the bulk as well as from regions near the surface. As a result, surface roughness has less effect than it has for metals. For semiconductors, the situation lies between that of a metal and an insulator, and complex phenomena can arise, which will be discussed in more detail below.

11.2.4 Optical Properties of Materials Present in Wafers

In order to predict the thermal radiative properties of semiconductor wafers, it is necessary to have a good knowledge of the optical properties of the materials involved. Timans conducted an extensive review of the available data for many of the materials involved in device fabrication, and only some key results for the most common materials encountered in silicon processing will be discussed here [14].

11.2.4.1 Optical Properties of Silicon

Data for the room temperature refractive index and extinction coefficient as a function of wavelength can be found in several recent reviews [20,21]. In RTP one is typically interested in the behavior at high temperatures, where less data are available. Table 11.2 summarizes some key studies of optical properties of silicon, categorizing them according to the properties measured and method used, wavelength range, temperature, and the state of silicon including doping and microstructure [20–36].

TABLE 11.2 References for Optical Properties of Various Forms of Silicon

Form	Doping	Property	λ Range (μm)	T Range (°C)	Measurement Method	Comment	Reference
c-Si	—	n and k	0.0006–333	RT	—	Review of room temperature properties	[20]
c-Si	—	n, k and α	0.003–333	RT	—	Review of room temperature properties	[21]
Various	Various	n and α	0.4–20	RT-800	—	Review of optical properties of semiconductors at elevated temperatures, includes models for n and α	[14]
c-Si	—	n, k and α	0.24–0.84	RT-490	Spectroscopic ellipsometry	Provides empirical fits, some of which are useful beyond range of measurements	[22]
c-Si	90 Ωcm n-type	α	1.2–9	345–723	Emission spectroscopy	Includes general model for infra-red (IR) absorption across a broad range of wavelengths and temperatures	[23]
c-Si	77 and 0.012 Ωcm n-type	α	1.15–1.6	343–758	Emission spectroscopy	Provides models for absorption in lightly doped silicon	[24]
c-Si	"very pure"	α	0.95–1.3	−269–142	Transmission	Detailed model for shape of absorption edge	[25]
c-Si	100 Ωcm	α	1.152	20–867	Transmission	Provides model for absorption	[26]
c-Si	Various	α	1.3 and 1.55	450–880	Transmission	Provides model for band-edge and free-carrier absorption	[27]
c-Si	>10 Ωcm n and p-type	α	1.7 and 3.4	400–700	Emissivity	Provides model for free-carrier absorption	[28]
c-Si	1.1×10^{15} cm^{-3} P-doped	α	9.5–11	25–500	Transmission		[29]
c-Si	15 Ωcm n-type	α	3–9	320–420	Emission spectroscopy	Describes the weak temperature dependence of lattice absorption	[30]
c-Si	Various	α	4–15	160	Emission spectroscopy	Lattice absorption study	[31]
c-Si	"High-purity"	α	4.5–10	22–167	Transmission	Lattice absorption study	[32]
c-Si	Various	N	1.2–14	−173–477		Review of data for $n(\lambda,T)$. Provides empirical fit	[33]
c-Si	Various, lightly doped	N	0.6–10	27–427	Interference measurements	Provides empirical fit for $n(\lambda,T)$	[34]
c-Si	Various doped	n, k and α	~2–50	RT		Summary of data for doped silicon	[35]
c-Si	Various doped	n and α	~0.4–~500	RT		Review of various data for doped silicon	[36]
p-Si and a-Si	Various	n and α	Various	Various		Review of properties and their temperature dependences	[14]

n, refractive index; k, extinction coefficient; α, absorption coefficient; λ, wavelength; T, temperature; and RT, room temperature.

Figure 11.5 shows absorption spectra of lightly doped silicon for a range of temperatures and wavelengths of interest for RTP technology [23]. The spectra show features produced by various absorption mechanisms. At short wavelengths, the absorption coefficient is very large, because the photons have enough energy to create electron–hole pairs. As the wavelength increases, the photon energy decreases and when it drops below the indirect silicon band-gap there is a very rapid decrease in α, called the absorption edge. Beyond the edge, the absorption coefficient is very small at low temperatures. Free charge carriers introduced by doping or by the finite temperature cause absorption which rises gradually with wavelength and dominates the behavior beyond the edge. This free-carrier absorption increases rapidly as the temperature rises and the concentration of thermally generated electron and holes increases. At low temperatures, there are weak absorption features superimposed on the free-carrier background. These features arise from absorption introduced by the vibrating ions of the silicon lattice, and they are swamped by the free-carrier effects once the temperature exceeds $\sim$400°C.

Figure 11.5 includes predictions from a semi-empirical model which can be used to calculate the absorption coefficient of lightly doped silicon in the infra-red (IR) at wavelengths between $\sim$0.9 and 9 μm. The model includes elements to describe the absorption related to interband transitions, free-carrier absorption, and lattice absorption. It gives good agreement with the experimental results for wavelengths between 1 and 9 μm and temperatures between 340 and 720°C. The model provides a simple method for understanding the optics of semi-transparent silicon slabs, as will be shown below.

In order to perform calculations of the spectral emissivity, we also need to know the wavelength and temperature dependences of the real part of the refractive index of silicon. Table 11.2 includes references to some useful results and models [22,33,34]. Figure 11.6 summarizes the behavior. The temperature and wavelength dependences are quite weak and can be ignored in many situations.

The effect of doping on the optical properties of silicon has been studied extensively [35,36]. Doping introduces free-carrier absorption, and its qualitative behavior can be predicted from the semi-classical Drude model that has been extensively applied to metals and semiconductors [14]. It assumes that the absorption arises from damping forces on the motion of charge carriers in an oscillating electric field. The model predicts that the free-carrier absorption coefficient rises with the square of the wavelength and that there is an effect on the refractive index of silicon. A more thorough analysis, including quantum-mechanical aspects, leads to different results, although the general trends are similar. Figure 11.7 shows the effect of doping on the absorption spectrum of silicon at room temperature [35]. At long

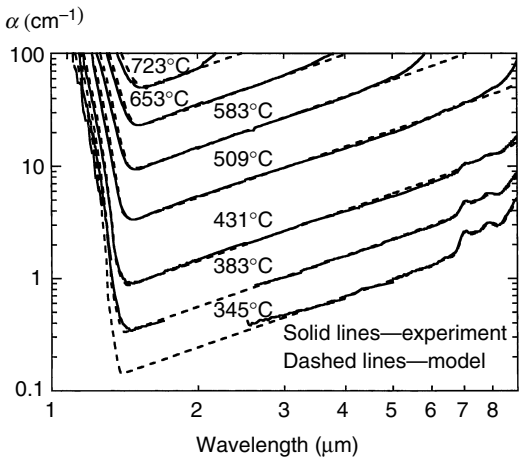


FIGURE 11.5 The absorption spectra of silicon for a range of temperatures. (From Rogne, H., Timans P. J., and Ahmed, H., *Appl. Phys. Lett.*, 69, 2190, 1996.)

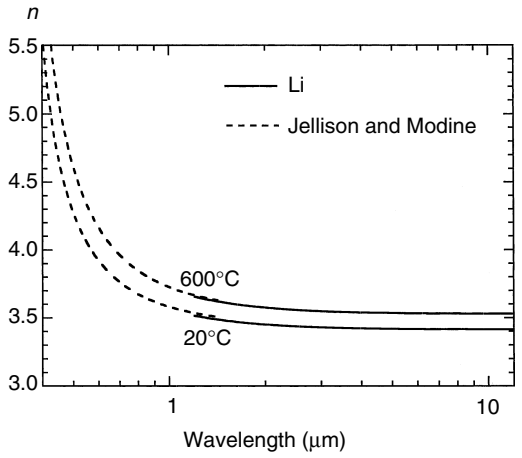


FIGURE 11.6 The refractive index of silicon, calculated from expressions given by Li (From Li, H. H., *J. Phys. Chem. Ref. Data*, 9, 561, 1980.) and by Jellison (From Jellison, G. E. Jr., and Modine, F. A., *J. Appl. Phys.*, 76, 3758, 1994.)

wavelengths, holes introduce more absorption than would arise from an equivalent concentration of electrons. Other effects, including additional absorption bands and changes in the shape of the fundamental absorption edge, can also occur through doping.

The crystal microstructure also influences the optical properties, especially for thin films of polycrystalline silicon (polysilicon) or amorphous silicon [14]. The differences in the optical properties of polysilicon and single crystal silicon are mainly caused by the grain boundaries. The grain boundaries disrupt the long-range crystalline order, which affects the shape of the absorption edge, where band-structure-related effects are prominent. At longer wavelengths, beyond $\sim 1.3\ \mu m$, the behavior for polysilicon films is similar to that of single crystal silicon. However, polysilicon films can have rough surfaces that can affect light scattering and thin-film interference effects in these films. The optical

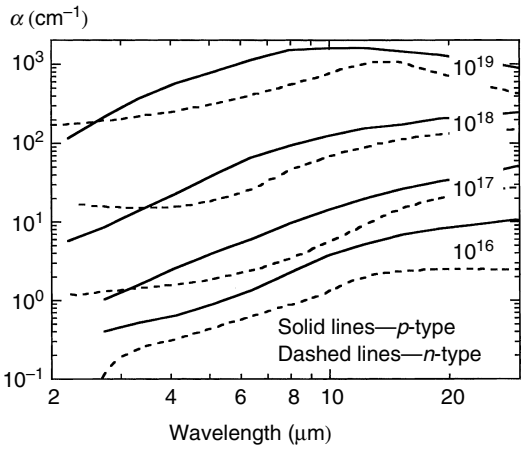


FIGURE 11.7 The absorption spectra of silicon for various concentrations of *p*- and *n*-type doping (From Schumann, P. A. Jr., Keenan, W. A., Tong, A. H., Gegenwarth H. H., and Schneider, C. P., *J. Electrochem. Soc.*, 118, 145, 1971). The concentrations are labeled in cm^{-3}.

properties of amorphous silicon are strongly affected by the deposition method because of differences in density, impurities, and defect structure. In particular, the passivation of dangling bonds by hydrogen alters the optical properties. The electronic structure is also radically different from that of crystalline silicon and, as for polysilicon, the largest differences in optical properties occur at wavelengths below 1.1 μm, but they are still significant even at longer wavelengths.

11.2.4.2 Optical Properties of Dielectrics

Table 11.3 summarizes some of the most useful references to studies of optical properties of dielectrics [14,37,38]. The optical properties of silicon dioxide, SiO_2, films are often assumed to be the same as those for bulk fused silica, whose optical properties have been studied extensively, at least at room temperature. The optical properties of SiO_2 films formed by thermal oxidation are very similar to those of bulk SiO_2. Some deposition methods result in films with significantly different densities and stoichiometries. Impurities can also alter the optical properties, especially in the IR because of their effect on lattice vibrations. Figure 11.8 shows the refractive index and absorption coefficients of bulk SiO_2. SiO_2 is transparent at wavelengths between 0.3 and 2.0 μm. Beyond 2.0 μm, absorption associated with impurities and the vibrations of the Si–O bond dominate the optical properties. The temperature dependence of the optical properties is weak, and can be ignored when modeling the optical properties of thin layers of SiO_2 on wafers.

Silicon nitride is formed by a number of processes including thermal growth, chemical vapor deposition (CVD), and sputtering. Its stoichiometry varies widely, depending on the means of preparation, and its optical properties have received little experimental attention. In summary, it is transparent for wavelengths in the visible and near IR, and has a refractive index $\sim$2.0. At wavelengths beyond $\sim$6 μm, lattice absorption dominates the behavior.

11.2.4.3 Optical Properties of Metals and Silicides

The optical properties of metals are very different to those of semiconductors and dielectrics. The very large concentration of free electrons makes free-carrier absorption dominate the optical properties, and the Drude model provides a reasonable explanation of the optical properties in the IR. This model can be used to predict the high-temperature optical properties in the IR from the temperature dependence of electrical properties, which is often described in the literature. At shorter wavelengths, including the visible spectrum, interband absorption effects cause large deviations from the Drude behavior. Metals exhibit a high reflectivity and a large absorption coefficient, which cause radiation to be absorbed very near the surface. This makes the surface condition very important and theoretical predictions of the optical properties do not always match experimental measurements closely because surface imperfections including roughness have large effects. Table 11.4 includes references to the optical properties of metals and metal compounds encountered in silicon device manufacturing [39–53]. Since metal silicides are generally good conductors, their behavior is rather similar to that of metals.

TABLE 11.3 References for Optical Properties of Dielectrics

Material	Property	λ Range (μm)	T Range (°C)	Comment	Reference
SiO_2	n, k and α	0.05–500	RT	Review with table of values	[37]
Si_3N_4	n, k and α	0.05–1.24	RT	Review with table of values	[38]
SiO_2 and Si_3N_4	n and α	Various	Various	Review including discussion of temperature dependence of properties	[14]

n, Refractive index; k, Extinction coefficient; α, Absorption coefficient; λ, Wavelength; T, Temperature; and RT, Room Temperature.

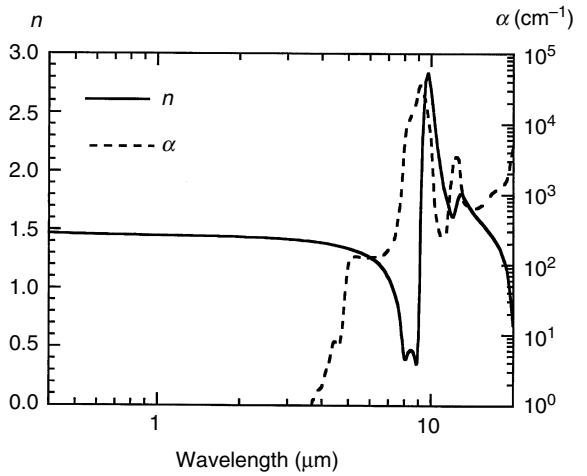


FIGURE 11.8 The refractive index, n, and absorption coefficient, α, of silicon dioxide (From Philipp, H. R., in *Properties of Silicon*, 1015–27 INSPEC, The Institution of Electrical Engineers, London, 1988.)

11.2.5 Thermal Radiative Properties of Semiconductor Wafers

Now that we have reviewed both the techniques for calculation of the thermal radiative properties and the fundamental optical properties of semiconductors and the other materials involved in device fabrication, we can examine the experimental and theoretical data for the spectral and total emissivities of semiconductor wafers. We shall start by examining the case of uncoated wafers with smooth surfaces, and then consider the impact of coatings and surface roughness.

11.2.5.1 Spectral Emissivity of Silicon

There have been several studies of the thermal radiation emitted by silicon at elevated temperatures. Since the spectral emissivity of silicon depends on the sample thickness, doping, and surface conditions, as well as the experimental factors discussed in Section 11.2.2.2, the results of these experiments must be interpreted with considerable care, before any general conclusions about the spectral emissivity of "silicon" can be made. Table 11.5 reviews some studies of the thermal radiative properties of various samples of silicon [14,24,28,30,54–57].

It is of practical interest to predict the spectral emissivity of a slab of lightly doped silicon which is 725 μm thick, since this is typical of a 200-mm wafer. Figure 11.9 shows a prediction obtained by using the model for absorption spectra in Figure 11.5, combined with a predicted refractive index. This kind of calculation has shown to give good agreement with data from experimental studies. The emissivity is high at short wavelengths where the wafer is opaque, and exhibits a large decrease when the wavelength lies beyond the absorption edge, where the wafer is semi-transparent and the emissivity is dominated by the absorption spectrum of silicon. The calculations show that the wafer only becomes opaque at all wavelengths once the temperature is above 700°C. For short wavelengths, where the sample is opaque, the spectral emissivity decreases slightly as the temperature rises because the surface reflectivity increases with temperature. Figure 11.10 shows the effect of wafer thickness at several temperatures. Above ~800°C, the thickness makes no difference because the wafers are all opaque at all wavelengths of interest.

For heavily doped wafers, which are opaque, the behavior is quite different. Figure 11.11 shows a comparison between experimental data for a heavily doped wafer and a prediction based on the Drude model discussed above [55]. There is good agreement with calculations, which predict a minimum in the reflectivity of a heavily doped sample, as a result of a phenomenon known as plasma resonance. This minimum in reflectivity corresponds to a maximum in emissivity.

TABLE 11.4 References for Optical Properties of Conductors

Material	Property	λ Range (μm)	T Range (°C)	Measurement Method	Comment	Reference
Al	n, k and R	0.017–32	Mainly RT		Review of properties, including temperature dependence	[39]
Al	n and k	0.4–32	RT		Summary of studies	[40]
Co	n and k	0.0006–17	RT		Describes behavior for various forms of Co	[41]
Co	n and k	0.188–20	RT		Summary of studies	[40]
Co	n and k	0.38–3.39	20–1600	Beatte method	Study of temperature dependence of optical properties	[42]
Cu	n and k	0.00014–9.5	RT		Review	[43]
Cu	n and k	0.4–20	RT		Summary of studies	[40]
Ti	n and k	0.82–20	RT		Summary of studies	[40]
Ti	n and k	0.4–10	RT	Null polarization method	Electrolytically polished Ti	[44]
W	n and k	0.0006–24.8	RT		Review	[45]
$CoSi_2$	$\varepsilon 1$, $\varepsilon 2$ and R	0.06–20	RT	Ellipsometry and reflectivity	Epitaxial $CoSi_2$ films	[46]
$CoSi_2$	$\varepsilon 1$ and $\varepsilon 2$	1–20	RT	Transmission spectra	Thin films grown by molecular beam epitaxy	[47]
$CoSi_2$	n and k	1.2–2.5	RT	Transmission spectra	Thin films grown by molecular beam epitaxy	[48]
$TiSi_2$	$\varepsilon 1$ and $\varepsilon 2$	0.06–2	RT	Reflection spectra	C54-phase single crystals	[49]
$TiSi_2$	n, k and R	0.16–12.4	RT	Reflection spectra	Single crystal	[50]
$TiSi_x$	n, k and R	0.4–0.7	RT	Ellipsometry	Various stages of transformation from deposited Ti film to $TiSi_2$ studied	[51]
Various silicides	$\varepsilon 1$, $\varepsilon 2$ and R	0.2–24	RT	Reflection spectra	Deposited and reacted films, $TiSi_2$, $TaSi_2$ and WSi_2	[52]
TiN	n and k	0.4–2	RT	Reflection and transmission spectra	Thin film TiN deposited by sputtering	[53]

n, Refractive index; k, Extinction coefficient; R, Reflectivity; ε_1, Real part of dielectric function; ε_2, Imaginary part of dielectric function; λ, Wavelength; T, Temperature; RT, Room Temperature.

TABLE 11.5 References for Thermal Radiative Properties of Plain Silicon Wafers

Property	Doping	Thickness (μm)	Surface Finish	λ Range (μm)	T Range (°C)	Measurement Method	Comment	Reference
Various	—	—	—	—	—	—	Review of thermal radiative properties of silicon wafers	[14]
$\varepsilon_n(\lambda)$	Lightly doped	725	Smooth	0.4–15	300–800	Calculation	Calculation for typical 200 mm wafer thickness, based on full optical model	[54]
$\varepsilon_n(\lambda)$	15 Ωcm P-doped	1770	Double-polished	0.4–15	270–800	Emission, reflection and transmission spectra	Includes discussion of T dependence of lattice absorption	[30]
$\varepsilon_n(\lambda)$	0.007 Ωcm P-doped	200	Double-polished	2–15	270–620	Emission spectra	Samples were opaque—includes model for effect of doping	[30]
$\varepsilon_n(\lambda)$	Various, heavily doped, *n*- and *p*-type	1600	Frontside polished	2–36	27–801	Emission and reflection spectra	Samples were probably opaque	[55]
$\varepsilon_n(\lambda)$	Lightly doped, *n*- and *p*-type	625, 675	Various	1.7 and 3.4	~300–700	Pyrometer measurements	Effect of wafer roughness studied, includes models for absorption and emissivity	[28]
$\varepsilon_{tot}(T)$	77 and 0.012 Ωcm, P-doped	390	Double-polished	—	340–800	Measurements from electron beam heating RTP system	Temperature dependence of total emissivity measured and compared to theory	[24,56]
$\varepsilon_{tot}(T)$ and $\alpha_{tot}(T)$	Lightly doped	725	Smooth	—	RT-800	Theoretical calculations	Calculated values for total normal and hemispherical properties. αtot calculated for lamps at 2000°C	[54]
Various integrated properties	Lightly doped	725	Smooth	—	RT-800	Theoretical calculations	Integrated properties calculated for bands between 0.4–4 μm and 4–25 μm. Relevant to heat transfer in quartz-wall RTP	[57]

λ, Wavelength; T, Temperature; RT, Room temperature; $\varepsilon_n(\lambda)$, Normal spectral emissivity; ε_{tot}, Total emissivity; α_{tot}, Total absorptivity.

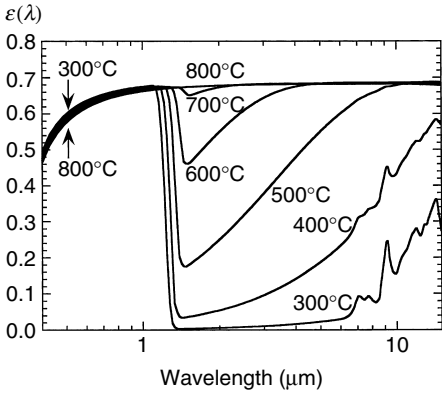


FIGURE 11.9 Predicted spectral emissivities for a 725-μm thick, lightly doped silicon wafer with smooth surfaces, for a range of temperatures. At temperatures above ~800°C, the wafer is completely opaque, and the spectral emissivity becomes a weak function of temperature. The spectra were calculated from a theoretical model (From Timans, P. J., in *Advances in Rapid Thermal and Integrated Processing*, edited by Roozeboom, F., 35–102, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.)

11.2.5.2 Integrated Emissivities and Absorptivities for Plain Silicon Wafers

The integrated properties also vary significantly with changes in sample thickness, doping, and the surface roughness of the wafer. Figure 11.12 illustrates experimental measurements of the total emissivities of 390 μm thick samples of lightly and heavily doped silicon, together with theoretical predictions of the total normal and total hemispherical emissivities [56]. The total emissivity of the lightly doped sample rises rapidly once the temperature exceeds ~400°C, because of the increasing free-carrier absorption. For the heavily doped sample there is a high concentration of free-carriers even at low temperature, and as a result the emissivity is high at low temperatures and it is not strongly

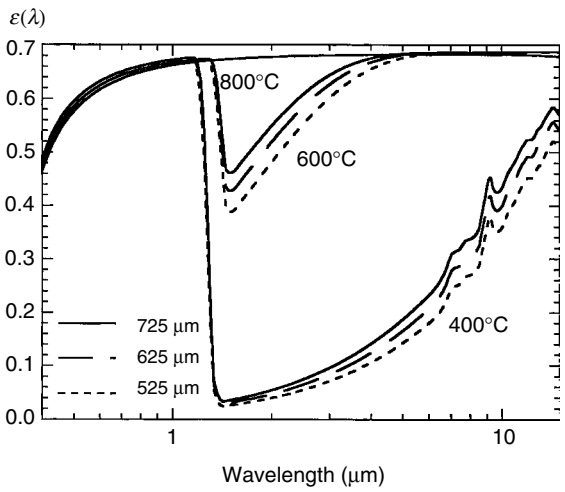


FIGURE 11.10 Predictions of the effect of silicon wafer thickness on spectral emissivity, for a few typical wafer thicknesses. The spectral emissivities were calculated from a model (From Timans, P. J., in *Advances in Rapid Thermal and Integrated Processing*, edited by Roozeboom, F., 35–102, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.)

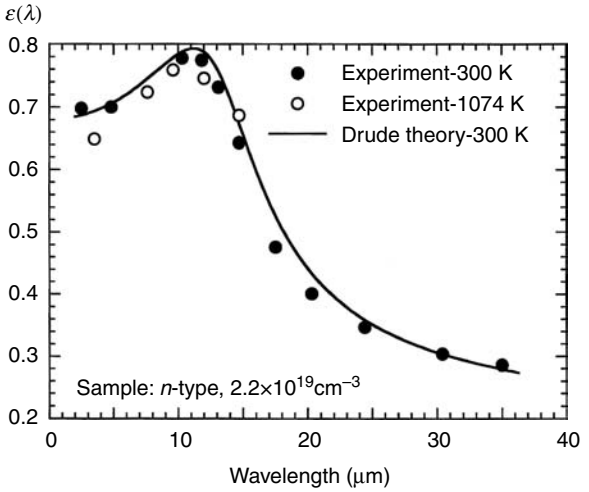


FIGURE 11.11 The spectral emissivity of a heavily-doped silicon surface. The solid line is obtained from a theoretical model which uses the Drude theory to describe the optical effect of heavy doping (From Timans, P. J., in *Advances in Rapid Thermal and Integrated Processing*, edited by Roozeboom, F., 35–102, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996). The points are from measurements by Liebert (From Liebert, C. H., *Progress in Astronautics and Aeronautics*, 20, 17–40, 1967.)

temperature dependent. The theoretical predictions of total emissivities were performed using the methods described above and assuming that the refractive index of silicon is 3.6. There is little difference between the predicted values of total normal and hemispherical emissivities.

Figure 11.13 shows predictions of total emissivities and absorptivities for a 725-μm thick wafer. The lamps were treated as blackbody sources at either 2000 or 2500°C, with a wavelength range from 0.4 to 4 μm. There is little difference between the total normal and hemispherical values of emissivity and absorptivity. The total absorptivity at room temperature is low and it depends on the lamp temperature.

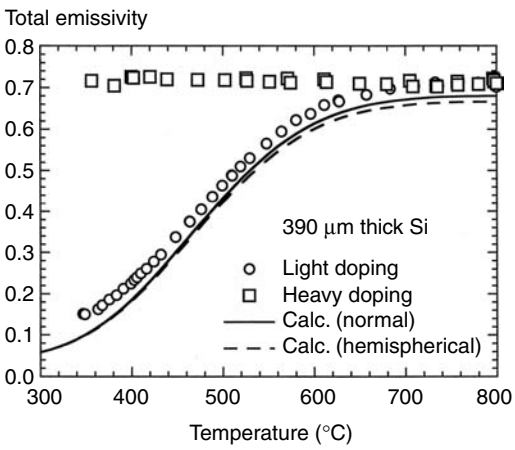


FIGURE 11.12 The total hemispherical emissivity of a 390-μm thick, lightly-doped silicon sample from measurements and from theoretical calculations (From Timans, P. J., in *Rapid Thermal Processing 1993*, edited by Fair, R. B. and Lojek, B., RTP 1993, Scottsdale, 1993, 282–6). The calculations were performed for both total normal emissivity and for total hemispherical emissivity.

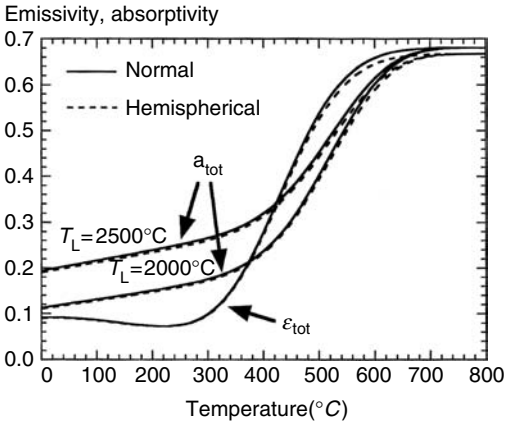


FIGURE 11.13 Predictions of the total emissivities (ε_{tot}) and absorptivities (a_{tot}) for a 725-μm thick, lightly-doped silicon wafer that has smooth surfaces. The predictions are shown for both normal and hemispherical cases. The absorptivities are shown for the cases where the lamps are assumed to be blackbodies at temperatures, T_L, of either 2000 or 2500°C.

The wafer is very transparent in the IR and lamp energy is absorbed only at $\lambda < \sim 1$ μm, where the wafer is opaque. For higher lamp temperatures, more of the lamp energy is at shorter wavelengths and as a result the wafer's absorptivity is higher. Once the wafer temperature exceeds ~400°C, the absorptivity rises rapidly because of the strong increase in the free-carrier concentration. The total emissivity is also low at room temperature when the emission spectrum is dominated by features associated with the weak lattice absorption bands at $\lambda > \sim 6$ μm. Above ~350°C, the total emissivity rises very rapidly with temperature and it soon exceeds the total absorptivity for lamp radiation, because a larger fraction of the wafer's emission spectrum lies at the long wavelengths where free-carrier absorption is strong.

11.2.5.3 Effects of Coatings on Thermal Radiative Properties

The changes in spectral emissivity produced by coatings on silicon are well-known to cause temperature errors in pyrometry, and much of the effort in studies of coating effects has been directed at this problem [61]. Table 11.6 summarizes some studies of the effect of coatings on the thermal radiative properties of wafers at elevated temperatures [14,54,58–77].

Coatings also affect the integrated emissivities and absorptivities of wafers. These influence the thermal response and variations in these properties over patterned areas of the wafer can even affect temperature uniformity on a wafer. This phenomenon will be discussed in more detail, below. Although thin-film coatings are the most obvious surface modifications which impact the thermal radiative properties of a wafer, doped layers can also have large effects. For example, during ion implantation damage annealing electrical activation of dopants can radically change the spectral and total emissivity of a wafer [77].

11.2.5.3.1 Examples of Behavior of Various Types of Wafer in RTP

There are myriad possibilities for what coating structures may be on a wafer, both for the frontside and for the backside, and it is not possible to discuss them all here, but is useful to assess the impact of some simple, common cases found on the backs of wafers, with the help of simulations. Table 11.7 describes the general impact of these structures on both the spectral emissivity and the integrated properties. This information can help assessment of the effects of these structures on temperature measurement, control, and uniformity optimization problems. Figure 11.14 through Figure 11.18 aim to provide a more detailed feel for the key properties of these structures, through representation of the spectral and integrated emissivities on graphs and contour maps.

TABLE 11.6 References for Thermal Radiative Properties of Coatings on Silicon Wafers

Types of Coatings	Property	Substrate Information	λ Range (μm)	T Range (°C)	Measurement Method	Comment	Reference
Various	—	—	—	—	—	Review of effects of coatings on spectra emissivities and integrated properties	[14, 54]
Various	—	—	—	—	—	Review of RTP physics includes effects of coatings on spectral emissivities and integrated properties	[58]
Thermal SiO_2, PETEOS, LPCVD TEOS, BPSG, a-Si-on-oxide, poly-on-oxide, Plasma Si_3N_4 and LPCVD Si_3N_4 on oxide	$\varepsilon(\lambda)$	—	0.9–9.2	750–1200	Emission spectra	Results for 45° angle-of-incidence, s-polarization. Results show good agreement with thin-film model. Optical properties of SiO_2 fitted for $\lambda > 6$ μm	[59]
Various SiO_2, Si_3N_4, poly-on-oxide, oxide-poly-oxide	$\varepsilon(\lambda)$	—	0.9–2.4	750–850	Emission spectra	Results for 45° angle-of-incidence, s-polarization. Good agreement with model	[60]
Various SiO_2	$\varepsilon(\lambda)$	0.015 Ωcm, B-doped	3.5	600–1100	Pyrometer and reflectivity measurement	Early study of coating effect on pyrometry, includes models for emissivity and reflective enhancement from chamber wall. Study included measurements of nitride-on-oxide and poly-on-oxide films	[61]
0.8 μm SiO_2 films	$\varepsilon(\lambda)$	10 Ωcm	4.5	250–1050	Pyrometer	Includes comparison of effects of smooth and rough surfaces	[62]
Various SiO_2 films, various poly films-on-0.5 μm oxide	$\varepsilon(\lambda)$	10 Ωcm	3.5 and 4.5	Mostly ~900	Pyrometers	Oxide films: 3000–8500 Å, Poly films: 1000–9000 Å. Unpolished surface coated for most samples. Pyrometer studies in reflecting chambers, includes model	[63]
SiO_2, Si_3N_4 and poly-on-oxide films	$\varepsilon(\lambda)$	—	3.3 and 4.5	800 and 1000	Pyrometers	Includes deduced values for optical constants of films and estimates of chamber reflectivities	[64]
0.1 and 1.3 μm SiO_2 films	$\varepsilon(\lambda)$	2–6 Ωcm, P-doped, <111>	9.4 and 8–14	100–500	Pyrometers	Study centred on Si–O absorption bands	[65]
0.51 μm SiO_2 film	$\varepsilon(\lambda)$	—	1.5–20	58–915	Emission spectra	Study using Fourier transform IR spectroscopic emissometer	[66]

(*continued*)

TABLE 11.6 ***(Continued)***

Types of Coatings	Property	Substrate Information	λ Range (μm)	T Range (°C)	Measurement Method	Comment	Reference
Si_3N_4 and various poly-on-oxide films	$\varepsilon(\lambda)$	—	0.95 and 1	~100–700	Ripple pyrometer	Includes comparison to theory	[67]
Si_3N_4	$\varepsilon(\lambda)$	—	~2.2	460–970	Pyrometer	Early study of effect of coatings on pyrometer error	[68]
Various SiO_2 films, Ti films on SiO_2 and Si	$\varepsilon(\lambda)$	—	2.6–3.5, 3.5 and 4.5	300–1100	Pyrometer	Includes observations of emissivity changes for Ti films during reactions	[69]
Al films	$\varepsilon(\lambda)$	—	0.8–1.1	20–580	Reflectivity	60° angle-of-incidence measurement	[70]
Co silicide	$\varepsilon(\lambda)$	Lightly doped silicon	0.6–3.2	400–800	Emissivity-corrected pyrometry	In situ measurements of Co silicide formation	[71]
Oxide films, <1 μm thick	a_{tot} and ε_{tot}	—	Various lamp and wafer spectra	1100	—	Analysis of pattern-induced temperature non-uniformities, includes chamber effects	[72]
Poly-on-oxide	a_{tot}	—	Arc lamp	~650	—	Study of interaction between deposition non-uniformity and lamp coupling	[73]
Oxide and polysilicon-on-oxide films	a_{tot} and ε_{tot}	—	W-halogen lamp and wafer spectra	RT-~800	—	Theoretical study of impact of coatings on properties and temperature uniformity during RTP and rapid thermal chemical vapor deposition (RTCVD)	[74,75]
Implanted layers	$\varepsilon(\lambda)$	Various substrates	3.4	RT-700	Pyrometer	Examined emissivity of both implanted and unimplanted surfaces of double-side polished wafers. Various doses of As, P and B ions	[76]
B-implanted layer	$\varepsilon(\lambda)$, ε_{tot}	Lightly doped substrate	1–2.7	300–800	Measurement in electron beam heating RTP system	Study of the effect of annealing a B-implanted layer (30 keV, 10^{16} B^+/cm^2) on both spectral and total emissivity	[77]

λ, Wavelength; T, Temperature; RT, Room temperature; $\varepsilon(\lambda)$, Spectral emissivity; ε_{tot}, Total emissivity; a_{tot}, Total absorptivity.

TABLE 11.7 Typical Effects of Various Generic Coatings on the Thermal Radiative Properties of Silicon

Type of Coating on Silicon	Effect on $\varepsilon(\lambda)$	Effect on Integrated Optical Properties	Comments
No coating	—	—	Spectrally grey with $\varepsilon(\lambda) \sim 0.7$
Amorphous Si	Small oscillations, maximum emissivity = emissivity of silicon	Small decrease in emissivity	Small effect, crystallization produces a change in properties
Undoped polysilicon ($<10^{18}$ cm^{-3})	Very small effect, effect is largest in near-IR	Negligible effect	Changes in surface roughness can affect the behavior
Films over undoped polysilicon directly on top of silicon	Behavior as for films above bare silicon	Behavior as for films above bare silicon	Since the polysilicon is much like the silicon beneath it, it has little effect
Doped polysilicon ($>10^{19}$ cm^{-3})	Complex free-carrier induced phenomena	Generally a fairly small effect	Film structures involving heavily doped silicon films can have properties which are difficult to predict
Thin oxide (<100 Å)	Negligible effect	No effect	Very little impact on properties
Medium oxide (100–2000 Å)	Increase in emissivity	Increase in lamp power absorption	
Thick oxide (>2000 Å)	Oscillations, minimum emissivity = emissivity of silicon, maximum emissivity ~0.92	Increased lamp power absorption and total emissivity	For oxide thickness >1 μm, the integrated optical properties are not very sensitive to thickness
Doped oxides, e.g., BPSG	Much like oxide, IR absorption from dopant-related bonds affects behavior at $\lambda > \sim 4$ μm	Very similar to oxide	
Thin nitride (<1000 Å)	Increase in emissivity	Increase in power absorption	
Nitride (>1000 Å)	Oscillations, minimum emissivity = emissivity of silicon, maximum emissivity ~1.0	Increased lamp power absorption and total emissivity	
Undoped poly on oxide (oxide <300 Å)	Oscillations in emissivity	Little effect	For oxide <100 Å, effects are small
Undoped poly on oxide (oxide >300 Å, poly <1500 Å)	Large oscillations in emissivity, typically between ~0.2 and 1.0	Strong decrease in lamp absorption and significant effect on total emissivity	
Undoped poly-on-oxide (oxide >300 Å, poly >1500 Å)	Large oscillations in emissivity, typically between ~0.2 and 1.0	Significant effects	As the oxide and polysilicon layers become thicker, effects on integrated properties become less critically sensitive to thicknesses
Structures involving amorphous silicon films	Behavior much like that of polysilicon, crystallization can change spectral emissivity significantly	Behavior much like that of polysilicon, crystallization can result in small changes	
Multiple poly-on-oxide stacks	Potentially very large oscillations in spectral emissivity, from ~0 to 1	Large effects	
Nitride-on-oxide	Oscillations in emissivity, emissivity can go above or below that of plain silicon	Some effect	For films where oxide and nitride are <~100 Å thick, effects are small
Metals and silicides	Large impact on properties, typically decrease in emissivity and non-grey spectral properties	Typically exhibit low absorptivity and emissivity as compared to opaque silicon	Behavior varies significantly. Reactions and phase changes can change the properties, especially for silicides. Surface roughness and micro-pattern effects can be important

The observations are expected to be valid for the case where the silicon substrate is opaque, for the spectral range between ~0.8 and 10 μm. Substrate doping and the presence of epi-layers could modify the behavior in some cases. λ, Wavelength and $\varepsilon(\lambda)$, Spectral emissivity.

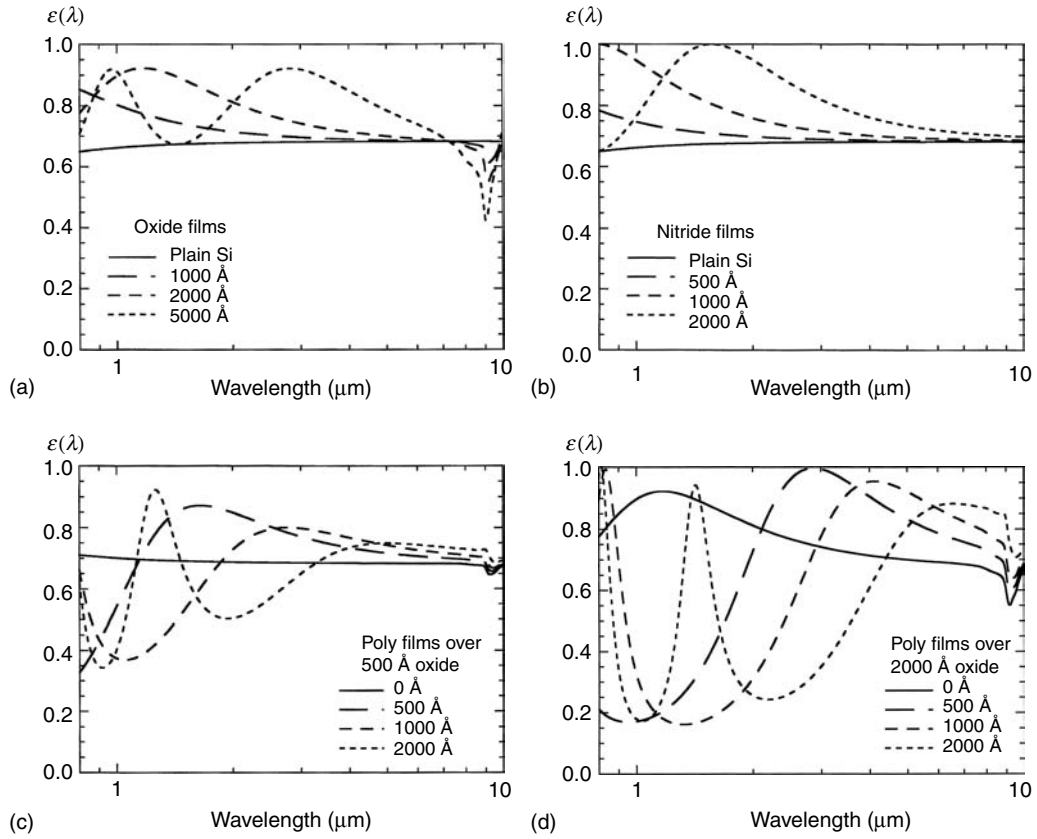


FIGURE 11.14 Predicted normal spectral emissivities for coatings on silicon. (a) Silicon dioxide films; (b) silicon nitride films; (c) polysilicon films on top of 500 Å of silicon dioxide; and (d) polysilicon films on top of 2000 Å of silicon dioxide.

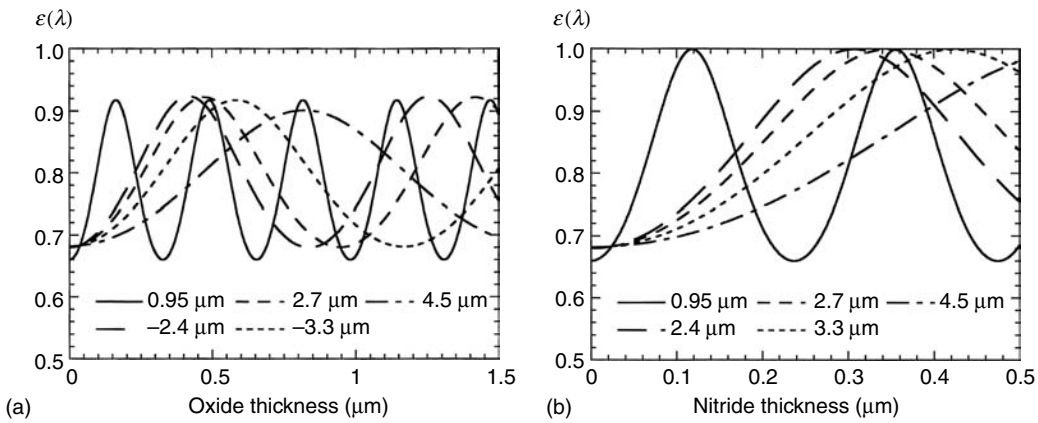


FIGURE 11.15 Predictions of the effect of coating thicknesses on the normal spectral emissivities of silicon at some typical pyrometer wavelengths. (a) Silicon dioxide thickness effect and (b) silicon nitride effect.

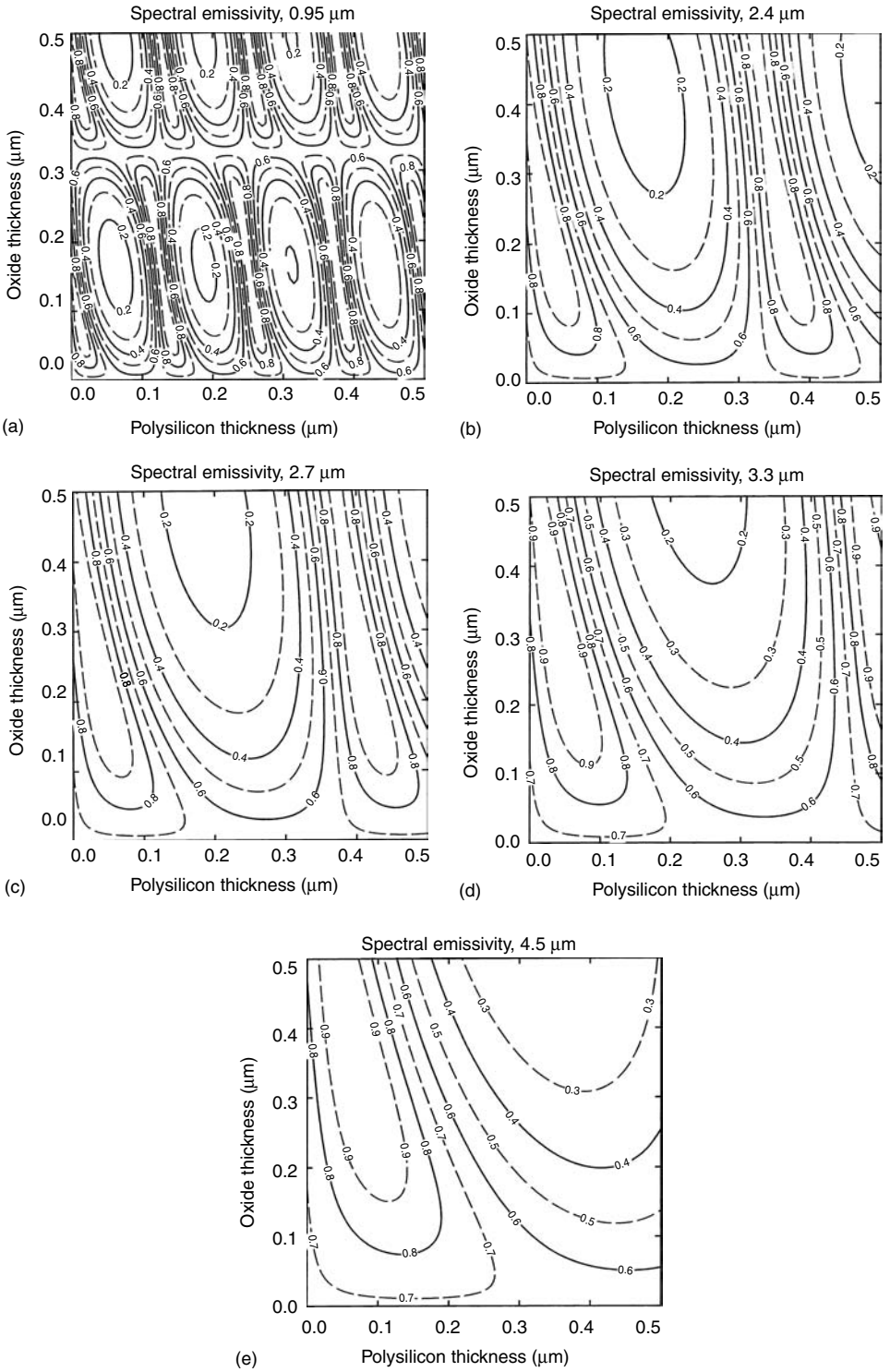


FIGURE 11.16 Contour maps showing predictions of the effects of silicon dioxide and polysilicon layer thicknesses on the normal spectral emissivity of silicon surfaces coated with polysilicon-on-oxide films. The results are for various pyrometer wavelengths (a) 0.95 μm; (b) 2.4 μm; (c) 2.7 μm; (d) 3.3 μm; and (e) 4.5 μm.

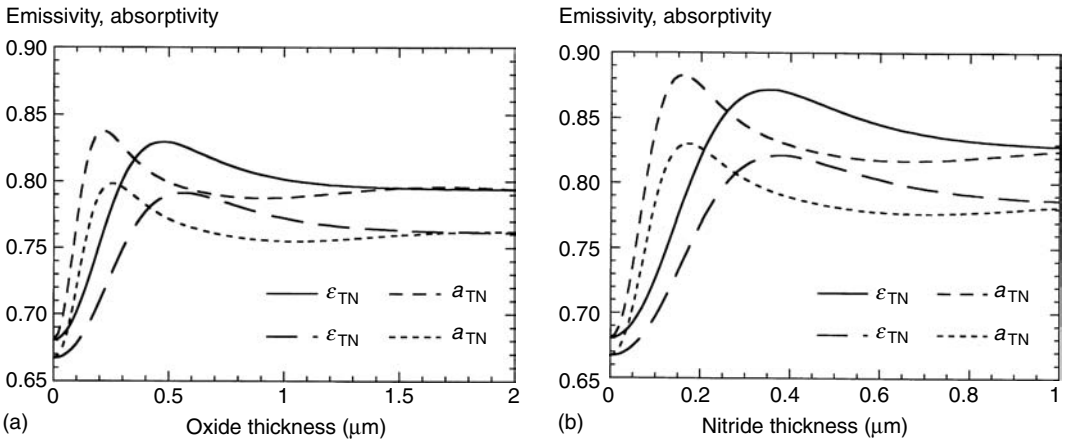


FIGURE 11.17 Predictions of the total emissivities, ε, and absorptivities, a, for silicon coated with (a) silicon dioxide and (b) silicon nitride films. The predictions are shown for both total normal (TN) and total hemispherical (TH) cases. The absorptivities were calculated with the assumption that the lamps are blackbodies at 2500°C.

Figure 11.14 shows the spectral emissivity for plain silicon, dielectric layers with several thicknesses, and several combinations of oxide layers with polysilicon layers. The latter are included because they tend to have large effects on optical properties. Figure 11.15 shows the effects of film thickness variations in oxide and nitride films on the spectral emissivites at wavelengths of 0.95, 2.4, 2.7, 3.3, and 4.5 μm, some of the more important wavelengths used in pyrometer control of RTP systems. Figure 11.16 shows similar data for polysilicon-on-oxide structures.

Figure 11.17 shows predictions for the normal and hemispherical integrated emissivities and absorptivities of oxide and nitride films as functions of their thickness. The lamps were treated as blackbody sources at 2500°C, with a wavelength range from 0.4 to 4 μm. The wafer was assumed to be at 1000°C. Figure 11.18 shows the effect of varying the thicknesses of the polysilicon and oxide layers in two-layer structures where the polysilicon film is on top of the oxide. In Figure 11.17 and Figure 11.18, a comparison of the predictions for the normal and the hemispherical properties indicates that the changes in coating thickness have slightly less effect on hemispherically integrated properties than on the normal properties. This is to be expected, because of the increased averaging of interference effects in thin-film coatings when multiple angles of incidence are taken into account.

11.2.5.4 Effects of Surface Roughness on Thermal Radiative Properties

The emissivity of an unpolished wafer surface can differ significantly from that of a smooth one. Some of these effects are strongly dependent on whether the substrate is opaque or not. At high temperatures, when a silicon substrate is opaque, surface roughness effects arise mainly from the change in the nature of the reflectivity of the surface and the effect on radiative properties is small unless the roughness is large enough for light to be multiply reflected within grooves on the surface [14,28,76,78]. Although the effect on emissivities is minor, roughness can still have a large impact on pyrometry, partly because of the interaction of light scattering effects with pyrometer system optics.

For semi-transparent wafers, the effect of roughness is quite different, because surface roughness modifies the absorption of radiation within the substrate by changing the nature of the internal multiple reflections illustrated in Figure 11.4. When radiation is incident on a sample with smooth surfaces from a given angle of incidence, all the rays within the substrate are at the same angle to the normal. If only the front surface is smooth, the rays internally reflected from the rough back surface acquire an angular distribution that depends on the nature of the roughness. Some of these scattered rays become trapped in the substrate by total internal reflection. The light-trapping greatly increases the average path length

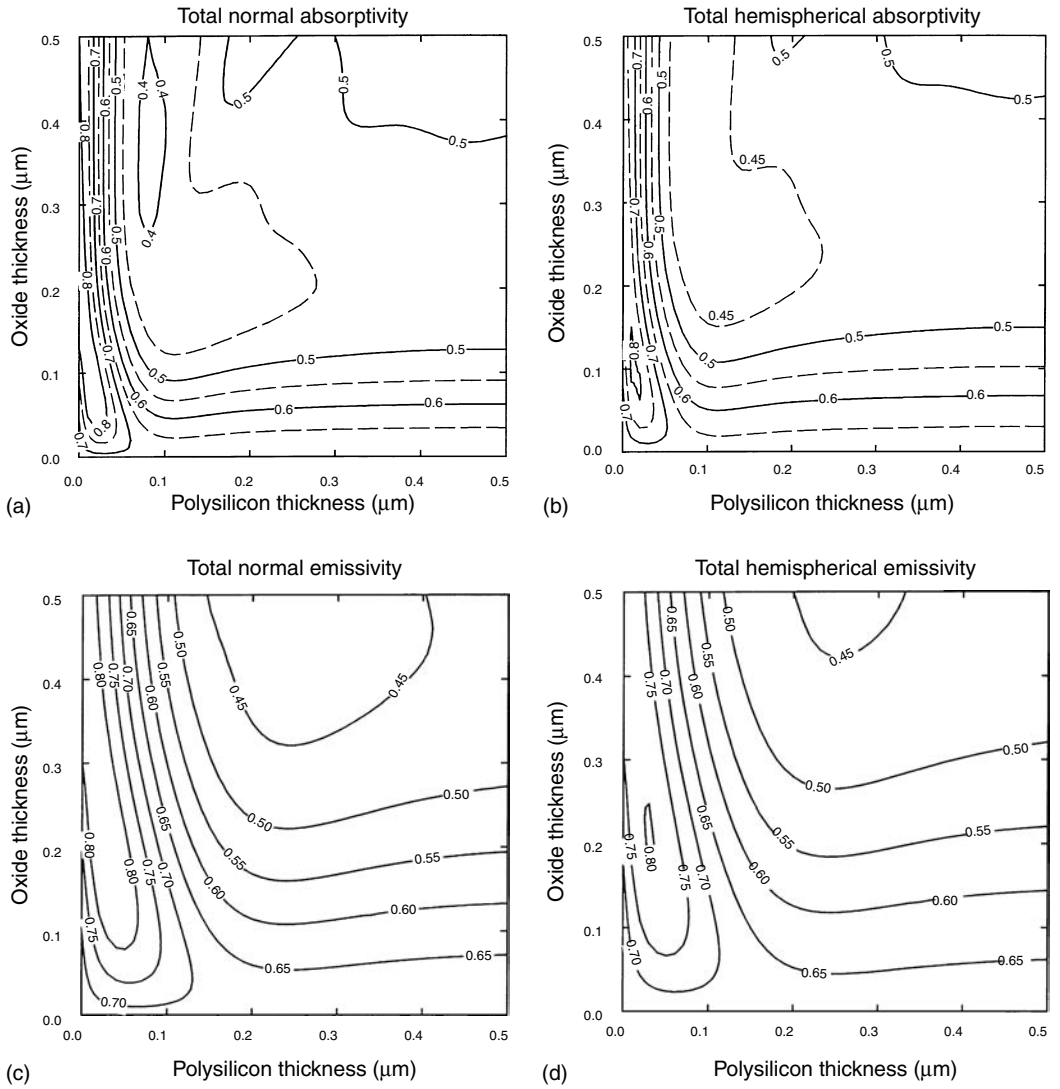


FIGURE 11.18 Contour maps showing predictions of the effects of silicon dioxide and polysilicon layer thicknesses on the total emissivities and absorptivities of silicon surfaces coated with polysilicon-on-oxide films. The results are for (a) normal absorptivity; (b) hemispherical absorptivity; (c) normal emissivity; and (d) hemispherical emissivity. The absorptivities were calculated with the assumption that the lamps are blackbodies at 2500°C and that the wafer is at 1000°C.

through the wafer and the fraction of radiation absorbed. This effect greatly increases the spectral and total emissivities and absorptivities as compared to those for similar samples with smooth surfaces [28].

11.2.6 Thermal Response of Wafers

11.2.6.1 Modeling and RTP

The physics needed to make accurate theoretical predictions of the thermal performance of RTP equipment is well understood. The difficulty in simulation of RTP systems lies mainly in the geometrical

and optical complexity of the problem, which requires models to describe features with large differences in length scales, from the shape of the process chamber itself to the details of lamp filaments. More complexity arises from the need for accurate descriptions of the radiative properties of the wafer and the chamber elements, including their spectral and surface scattering qualities. Nevertheless, the capability to perform simulation of the heating processes has allowed an excellent understanding of RTP equipment development requirements to evolve. Table 11.8 describes some theoretical studies where multi-dimensional computer models have illustrated important aspects of heat transfer, gas flow, and wafer stresses in RTP equipment [58,79–96]. Many important lessons about RTP system operation can also be learnt from simple models that capture the basic physics and provide guidance in system design and optimization [57]. In this section, we will use some simple models to consider a few of the basics of RTP heating.

11.2.6.2 Basic Heating Considerations

Much of the operation of RTP systems can be understood by examining one equation, which sums up the dynamic response of the wafer temperature to irradiation,

$$\rho c D \frac{\mathrm{d}T}{\mathrm{d}t} = \eta P(t) - H_{\mathrm{eff}} \sigma T^4, \tag{11.24}$$

where $P(t)$ is the total incident power density, t is time, T is the wafer temperature, η describes the power coupling efficiency, ρ and c are the density and specific heat capacity of silicon, respectively, D is the wafer thickness, σ is the Stefan–Boltzmann constant and H_{eff} is a parameter related to the integrated emissivity that describes the efficiency of power loss by thermal radiation [13]. Equation 11.24 reduces the 3D heat transfer problem to an equation which treats the wafer as a zero-dimensional object, since it does not allow for any temperature gradient within the wafer. It also ignores any convective or conductive heat losses through the gas ambient around the wafer, and the details of the radiation problem. The exact meanings of η, H_{eff}, and $P(t)$ depend on the nature of the RTP system and the wafer itself. The power may be incident from one side or both sides of the wafer, and H_{eff} includes the possibility of loss from both sides of the wafer, as well as re-reflection of emitted radiation back onto the wafer. $P(t)$ includes all sources of energy irradiating the wafer, not just the lamp power. For example, the chamber itself emits thermal radiation which tends to heat the wafer. This can become significant in warm-wall systems where quartz windows may be at temperatures above 400°C. In the steady-state, the equation reduces to the condition

$$\eta P = H_{\mathrm{eff}} \sigma T_{\mathrm{s}}^4, \tag{11.25}$$

where the power, P, assumes a fixed value corresponding to the steady-state temperature T_{s}. We will now use Equation 11.24 to deduce some characteristic features of RTP cycles.

11.2.6.2.1 Basic Thermal Response of Wafer to Lamp Radiation

Figure 11.19 shows the predicted thermal responses of a wafer to lamp radiation with a fixed total intensity for a period of 60 s. The intensities were chosen to result in peak temperatures between 600 and 1200°C. The wafer is assumed to have a fixed emissivity of 0.7 and to be in a black chamber. The calculation included the temperature dependence of the specific heat capacity of silicon and a fixed density of 2330 kgm^{-3}. As the steady-state temperature rises the time taken to settle at the steady-state condition reduces, and the temperature–time cycle looks more like the shape of the intensity cycle.

11.2.6.2.2 Cooling in RTP

When the power is switched off the temperature falls most rapidly initially, when the wafer is at the highest temperature. Figure 11.20 shows the predicted cooling rate as a function of temperature for a 725-μm thick silicon wafer cooling in a black cavity at 27°C. This is the fastest cooling rate that can be achieved through radiation cooling alone. In real systems, re-reflection of radiation onto the wafer,

TABLE 11.8 References to Multi-Dimensional Heat Transfer Simulations of RTP Phenomena

Type of System	Model Features	Key Phenomena	Experiments	Reference
Generic RTP	Simple 2D axisymmetric conduction (r and z in wafer), transient effects, spectral effects, includes wafer edge shape detail	Basic trends in uniformity, pattern effect	TC-wafer measurements	[58]
Generic RTP	Simple radiation exchange model, axisymmetric 1D conduction (r in wafer), transient effects, simple convective heat transfer, thermal stress	Basic trends in uniformity, effects of slip-free ring configuration, thermal stresses	Liquid crystal thermometry for low *T*, RTO tests for high T	[79]
Generic RTP	Simple 1D heat conduction (r in wafer), transient effects	Effect of illumination distributions on static and dynamic uniformity, pattern effect, slip-free ring	RTO and RTA tests of steady-state and transient effects	[80]
Axisymmetric	Radiation exchange model, axisymmetric 1D heat conduction (r in wafer), transient effects	Study of how to design axisymmetric lamp zone configuration for both static and dynamic uniformity	NA	[81]
Axisymmetric	2D heat conduction (in plane of wafer), ray-tracing analysis of radiation exchange	Study of the effect of the wafer edge effect	RTO tests	[82]
Photon box	Radiation exchange in 3D geometry, quartz tube "filtering" effect on spectrum	Effect of design features on uniformity: Patterned reflectors, crossed linear lamps, zone contouring, wafer rotation	TC wafer tests	[83]
AG 8108—photon box	3D radiation exchange model for a commercial RTP system with crossed linear lamp array, transient effects	Effect of zone power changes, slip-free ring effect, impact of patterned reflector surfaces	Multiple-TC wafer (17 TCs)	[84]
Axisymmetric	Multi-band radiation exchange model for axisymmetric system including details of quartz window effect, 1D heat conduction (r in wafer and quartz window), transient effects	Comparison of the effects of various assumptions about spectral radiative properties of chamber components	TC-wafer tests	[85]
Axisymmetric	Radiation exchange model using a 2-band approach to handle quartz window effect, transient effects, optimization of *T* uniformity, detailed gas flow calculations	Studies of lamp power requirements, gain profiles for controllability analysis, gas flow simulations	NA	[86]
Axisymmetric	Same model as in [86]	Simulated system identification for linear-quadratic-Gaussian MIMO control system, simulations of dynamic uniformity under MIMO control, effect of measurement disturbances on control	NA	[87]
Axisymmetric	2-band radiation exchange model, Monte-Carlo calculation of exchange factors, finite volume model for heat transfer, transient effects	Uniformity and power requirements, gain profiles, pattern effect, quartz window and showerhead heating effects	NA	[88]

(*continued*)

TABLE 11.8 ***(Continued)***

Type of System	Model Features	Key Phenomena	Experiments	Reference
Axisymmetric	Multi-band Monte Carlo model for radiation exchange including specular surfaces, 3D PHOENICS-CVD software model, transient effects	Effects of quartz window heat up on wafer *T* during open-loop processing	TC-wafer	[89]
Photon box	3D model similar to that in [89]. Detailed analysis of linear lamp radiation characteristics. Coupled calculations of heat transfer and gas flow patterns	Effect of gas flow on wafer *T*, quartz isolation tube heating, linear lamp radiation characteristics, gas flow patterns	RTO tests	[90]
Linear lamps + point source array	3D model similar to that in [89]. Includes detailed studies of lamp radiation from "point-source"-type W-halogen lamps	Study of simplified model, augmented by experimental data to make critical aspects very accurate. The simplified model is used as a process optimization aid and its efficacy is tested via the full 3D simulator	Comparison of measured and simulated radiation profiles from "point-sources"	[91]
Axisymmetric	2D heat transfer model	Experimental measurements of illumination profiles reveals the importance of accurate characterization of profiles through measurement	TC-wafers, illumination measurements	[92]
Photon box	Monte-Carlo ray trace for calculating heat transfer from lamps to wafer, heat conduction model for wafer	The model is used to decide how to change lamp zone parameters for process optimization. Method is robust with respect to process and hardware changes. Angular incidence of illumination, photon-box effect	RTO and RTA experiments	[93]
Axisymmetric	Full 3D analysis including electromagnetic field problem, radiation, heat conduction, gas flow, thermal stresses and chemical reactions, transient effects	3D analysis of effect of microscale patterns on radiative properties, pattern effect on *T* distribution and stresses, dynamic non-uniformity, gas flow effects on heat transfer, RTCVD deposition uniformity	NA	[94]
—	—	Review of RTP modeling	—	[95]
—	—	Review of RTCVD modeling	—	[96]

T, Temperature; TC, Thermocouple; *r*, radial co-ordinate; *z*, axial co-ordinate; RTO, Rapid thermal oxidation; RTA, Rapid thermal annealing; MIMO, Multiple-input, multiple-output; PHOENICS-CVD, Simulation package from Cham Ltd., London; and NA, Not available.

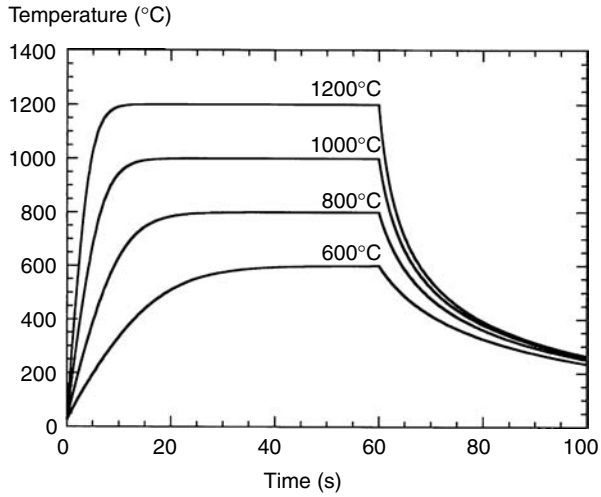


FIGURE 11.19 Predicted temperature cycles for heating a 725-μm thick plain silicon wafer with fixed power density for 60 s. The power densities were chosen to produce the steady-state temperatures shown.

and absorption of heat from the warm-wall environment would reduce the cooling rate. At temperatures below ~500°C, thermal conduction and convective heat losses can increase the cooling rate significantly.

11.2.6.2.3 Sensitivity to Fluctuations in Heat Transfer Conditions

Equation 11.25 can be used to determine the sensitivity of wafer temperature to fluctuations in η, H_{eff} and P [13]. For fluctuations $\Delta\eta$, ΔH_{eff} and ΔP, we can deduce the corresponding temperature fluctuation

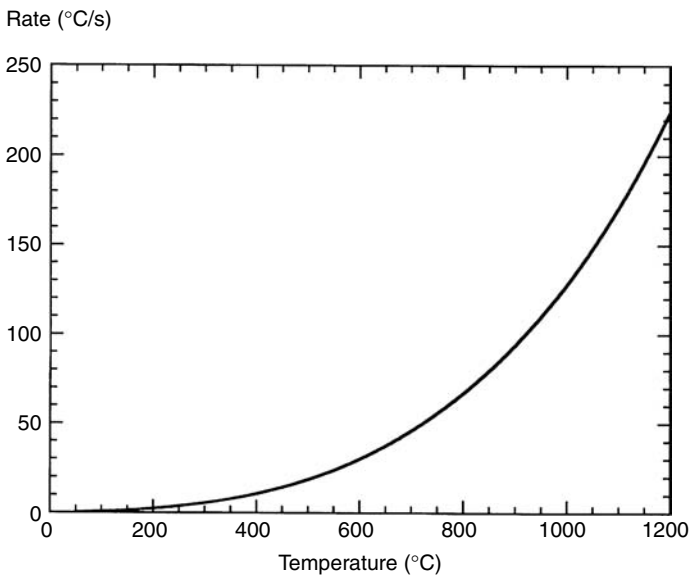


FIGURE 11.20 A prediction of the maximum radiation cooling rate for a 725-μm thick plain silicon wafer, shown as a function of its temperature. The calculation assumes that the wafer is in a perfectly black enclosure, which is kept at a temperature of 27°C.

ΔT,

$$\frac{\Delta T}{T} = \frac{1}{4}\left(\frac{\Delta P}{P} + \frac{\Delta \eta}{\eta} - \frac{\Delta H_{\text{eff}}}{H_{\text{eff}}}\right). \tag{11.26}$$

It is evident that variations in the incident power, power coupling, and effective emissivity are all equally significant in determining the temperature disturbance. Equation 11.26 predicts that a 1% change in the incident power produces a 0.25% change in the absolute temperature of the wafer. For example, if the wafer is at 1100°C, a 1% change in power density would change the wafer temperature by $0.0025 \times 1373\ K = 3.4$°C. This example illustrates the formidable technical challenge presented by the need to control the 3σ wafer temperature repeatability and uniformity to $\sim$1°C.

11.2.6.2.4 Power Requirements, Closed-Loop Control

The model can also be used to predict the power required to heat the wafer at a fixed ramp rate to a high temperature and then keep it there, as might happen when a wafer is heated under closed-loop control from a temperature sensor. Figure 11.21 illustrates the results for a few typical heating cycles. The power required gradually rises during the temperature ramp-up, as a result of the increasing radiation loss from the wafer. In a constant ramp-rate recipe, the peak power requirement occurs at the point immediately before the steady-state is reached, where the maximum radiation loss is combined with the nearly constant power needed to raise the wafer's temperature at the ramp rate.

11.2.6.2.5 RTP Peak Power Requirements

Equation 11.24 can be used to calculate the peak power requirements for an RTP system as a function of the process temperature and ramp rate required. Figure 11.22 is a contour plot illustrating the power requirements for various ramp rates and process temperatures, relative to the case for a 100°C/s ramp to 1000°C [13].

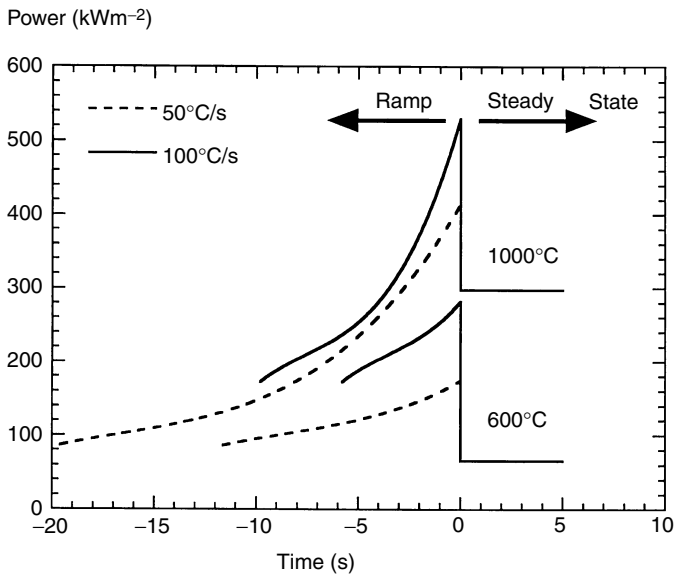


FIGURE 11.21 Predicted dynamic power density requirements for heating a 725-μm thick silicon wafer to either 600 or 1000°C at rates of 50°C/s or 100°C/s.

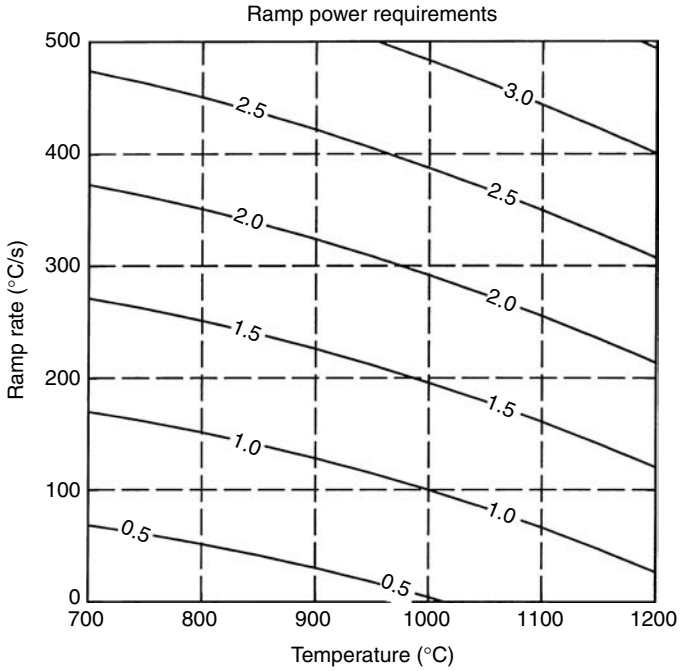


FIGURE 11.22 A contour map showing predictions of the peak power requirements for heating a wafer to various temperatures at various ramp rates. The results are normalized to the case for heating a wafer to 1000°C at 100°C/s.

11.2.6.3 RTP System Dynamics and Control Methods

Under closed-loop control, the control algorithm dynamically adjusts the lamp power settings to allow the wafer temperature to follow a prescribed processing recipe. There are many possible ways to implement the control, and this section merely explains some underlying factors in the system dynamics. The control problem is highly non-linear, because the wafer temperature is linked to the lamp power through a fourth-power law, but Equation 11.24 can be linearized to predict the dynamic change in wafer temperature, ΔT, caused by small changes in lamp power, ΔP, at a given temperature, T_0, to yield an equation of the form

$$\frac{\mathrm{d}(\Delta T)}{\mathrm{d}t} = \frac{\eta}{\rho c D}\Delta P - \frac{4H_{\mathrm{eff}}\sigma T_0^3}{\rho c D}\Delta T. \tag{11.27}$$

This equation can be used to form a first-order model of the dynamic response of the wafer, with a temperature-dependent gain, $G(T)$, and time constant, $\tau(T)$ [97]. The gain is given by

$$G(T) = \frac{\eta}{4H_{\mathrm{eff}}\sigma T^3} \tag{11.28}$$

and the time constant is given by

$$\tau(T) = \frac{\rho c D}{4H_{\mathrm{eff}}\sigma T^3}. \tag{11.29}$$

Figure 11.23 shows how these quantities vary with temperature, for a plain, 725 μm thick silicon wafer, where H_{eff} is assumed to equal 1.4. The gain, which describes how much the wafer temperature changes

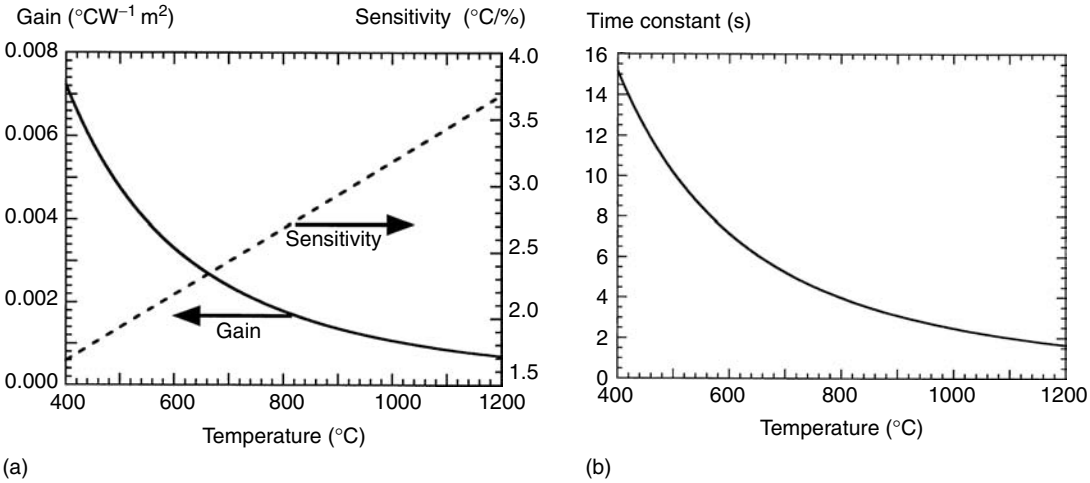


FIGURE 11.23 Predictions of the temperature dependences of parameters related to control of RTP process temperatures. (a) The gain and sensitivity of the wafer temperature. The gain describes the change in wafer temperature that results from a small change in lamp power density. The sensitivity describes the change in wafer temperature that arises from a given fractional change in the incident power. (b) The time constant for a wafer that is 725 μm thick.

for a given change in the lamp power density, decreases rapidly as the temperature rises. Another way of examining the impact of lamp power fluctuations is to consider the sensitivity to a given fractional change in the lamp power density. Figure 11.23a also includes the sensitivity of wafer temperature to a 1% change in lamp power density. This quantity rises linearly with the absolute temperature, because the heating power density rises as T^4, while the gain decreases with T^3. Figure 11.23b shows that the time constant of the wafer decreases as the wafer temperature rises. The time constant has an important impact on the system dynamics, including aspects such as the behavior of transient temperature non-uniformities.

The lamps also play an important part in the dynamics, and it is possible to construct a model for the lamp response which is similar to Equation 11.24. The lamp time constant is much smaller than that of the wafer, but it can still have an effect on the control problem [97].

Many types of control algorithm have been tested in RTP systems, including the well-known proportional-derivative-integral algorithm, as well as methods based on more complex approaches [98]. The parameters in the control algorithms can be established by trial-and-error approaches by the system user, who adjusts control parameters to achieve temperature–time responses without excessive overshoot or undershoot, as illustrated in Figure 11.24. More sophisticated systems use model-based controllers, where the parameters are obtained from empirical or physical models for the system response. The feedback–control algorithm is often combined with a feed-forward element that predicts the power requirements from a predetermined model of the system response and reduces the magnitude of the control action required by the feedback loop. Recent innovations in RTP control include the use of multiple temperature sensors to observe the temperature at several positions on the wafer to provide dynamic wafer temperature uniformity control [6–8,99]. This technique, called multiple-input, multiple-output (MIMO) control will be discussed further below.

11.2.6.4 A Simple Model for Heat Transfer in a Classic RTP Configuration

Many aspects of the operation of RTP systems, and the influence of the wafer's thermal radiative properties on the heating cycle can be assessed through the use of slightly more sophisticated heat transfer models. This section will briefly review a simple model for the common RTP configuration

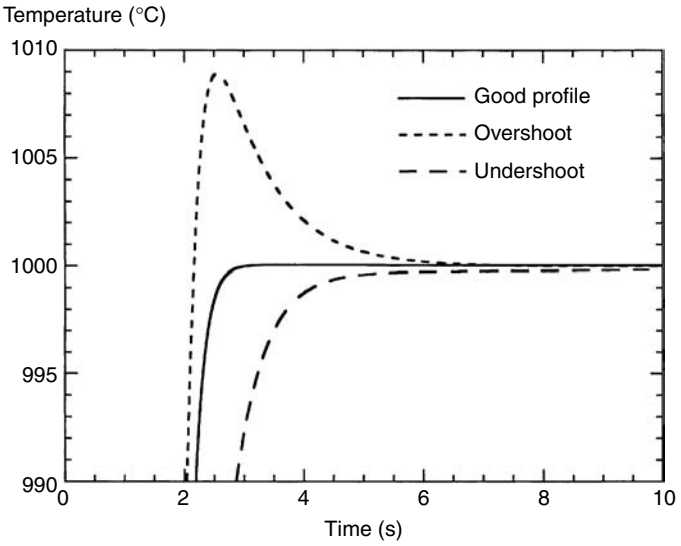


FIGURE 11.24 Control optimization.

shown in Figure 11.1a, and will use this model to derive several generic results of practical interest in RTP technology. The mathematical details of the model can be found in Ref. [57].

Figure 11.25 shows a simplified model for heat transfer in an RTP system in which the wafer is kept in a quartz isolation tube. There are lamps above and below the wafer, and the chamber has highly reflecting walls. The quartz tube is cooled by forced air, and the wafer is kept in a process gas ambient during heating. A zero-dimensional model can be constructed by representing the surfaces of the objects in the chamber as infinite parallel planes. These include the chamber walls, the two quartz plates of the isolation tube, and the wafer, each of which is assumed to be at one temperature. The geometry of the lamps is not included in the model, and their influence is modeled by assuming that the chamber walls emit uniform light fluxes.

Even in this simple model, the solution of the heat transfer problem is complex because the thermal radiative properties of the elements of the system vary with wavelength and temperature. In particular, the fact that the quartz tube is opaque at wavelengths beyond 4 μm, and transparent at shorter wavelengths, has a large impact on the behavior. It is natural to split the radiation problem into two wavebands, one for $\lambda < 4$ μm where the quartz is completely transparent and another for $\lambda > 4$ μm where it is opaque. The analysis is performed by separately considering the fluxes of radiation emitted by each energy source in each band.

The radiant energy sources in the system are the tungsten–halogen lamps, the wafer and the top and bottom plates of the quartz tube. The thermal radiation emitted by the chamber walls, which are usually water cooled, is neglected. The lamps are assumed to emit a blackbody spectrum at wavelengths from 0.4 to 4 μm, with a source temperature of 2500°C. For calculations of integrated thermal radiative properties, the quartz tube is assumed to be at a fixed temperature of 400°C, but the solutions allow this temperature to vary during heating cycles. The tube is assumed to radiate at wavelengths from 4 to 25 μm. Since the wafer temperature varies widely during the heating cycle, and its fundamental optical properties can also be strong functions of temperature, the integrated radiative properties of both surfaces of the wafer must be calculated as functions of its temperature, for each waveband and energy source included in the model.

The radiant power transfer between the various energy sources and objects in the system is deduced by analyzing how the energy emitted from each source flows between the various surfaces in the system. The set of equations for these energy fluxes is solved to determine the net radiation power input to each object

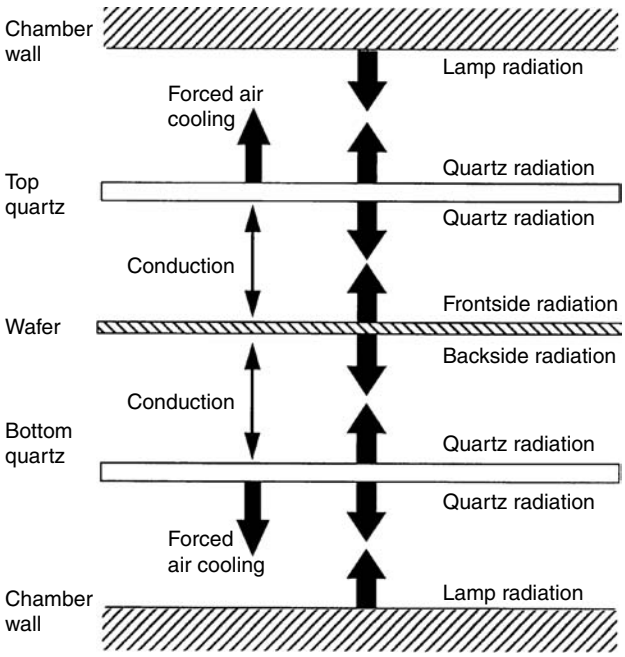


FIGURE 11.25 Simple model for heat transfer in a double-side-heated RTP system where the wafer is kept inside a quartz isolation tube (From Timans, P. J., in *Rapid Thermal Processing 1993*, edited by Fair R. B. and Lojek, B., 282–6; RTP '93, Scottsdale, 1993, Timans, P. J., *Solid State Technol.*, 40, (1997): 63.)

in the system. The model also accounts for the forced air-cooling of the quartz tube and includes a crude estimate of the conductive heat transfer via the process gas inside the tube. The analysis leads to a set of coupled differential equations which can be solved numerically to calculate the thermal response of the system when a given power is applied to the lamps. This model can be used for many purposes, and it is interesting to consider a few examples of what can be learnt about RTP system behavior from this simple simulation tool.

11.2.6.4.1 Effect of Wafer Substrate Doping on System Response

One interesting question is that of the transient response of different types of wafer under open-loop control. Figure 11.26 shows calculated temperature–time cycles for lightly and heavily doped silicon wafers when they are exposed to a fixed total intensity of either 40 or 100 kWm^{-2} that is divided equally between the lamps above and below the wafer. Figure 11.26 shows that the wafers reach the same steady-state temperatures but their transient responses differ, because the lightly doped material is semi-transparent at temperatures below 700°C and consequently it couples weakly with the lamp radiation and heats up slower. The difference in the heating rates of lightly and heavily doped wafers has been observed experimentally [64].

11.2.6.4.2 Effect of Quartz Temperature on Heating Cycles

Figure 11.27 shows the effect of multiple heating cycles on an open-loop RTP system response, for two values of the heat transfer coefficient for the air cooling of the quartz, h_q, 30 and 60 $Wm^{-2}K^{-1}$. The model illustrates the effect of quartz heating up. The wafer temperature at the end of each heating cycle gradually climbs as more heating cycles are performed. This situation, which could arise under open-loop control conditions, could introduce a repeatability problem in processing [89].

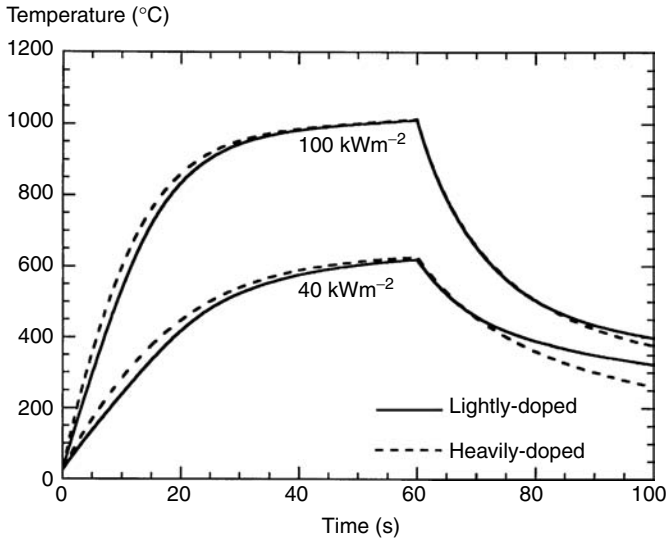


FIGURE 11.26 Prediction of the effect of wafer doping on thermal response of wafers during fixed intensity heating in a system like that shown in Figure 11.25. The numbers refer to the total power density used to heat the wafer.

11.2.6.5 More Sophisticated Models for Heat Transfer in RTP

In order to address the more complex aspects of heat transfer in RTP systems, in particular calculation of the temperature distribution across the wafer, one needs more sophisticated models for heat transfer, which include details of the system geometry. The general solution for a 3D geometry with heat

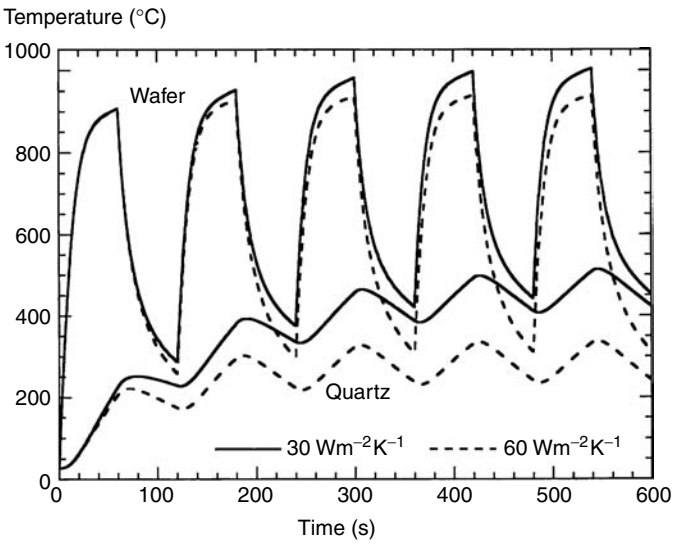


FIGURE 11.27 Illustration of the effects of quartz heat up on open-loop operation in a system like that shown in Figure 11.25. The calculations were performed assuming that the wafer is left inside the chamber and heated repeatedly with 60 s-duration fixed intensity blocks. Results are shown for two values of the heat transfer coefficient that describes the forced air cooling of the quartz tube. For the higher cooling rate the quartz temperature stabilizes faster, and the wafer heating cycles are more repeatable.

transfer via radiation, conduction, and convection is extremely complex, and requires finite-element or finite-volume heat transfer and fluid dynamics models which are very time consuming to construct and use. Table 11.8 refers to studies that used various types of model to analyze several aspects of the problem. Some studies include analysis of stresses induced in the wafer by thermal non-uniformities, which can have important implications for the generation of defects [79,94]. For analysis of RTCVD systems, all the heat transfer phenomena of RTP have to be included, but it is also important to include the role of chemical reactions in the model, which greatly increases the complexity [96].

11.2.7 Temperature Measurement and Control in RTP

11.2.7.1 Temperature Control Problems in RTP

The ± 2°C, 3-sigma specification for RTP temperature control deduced from the semiconductor industry association's (SIA) roadmap results in two fundamental technical problems. The first is that of making sure that the temperature distribution within any given wafer processed in an RTP system is well within this limit. This "wafer uniformity" problem will be discussed in a later section. The second problem is that of ensuring that every thermal cycle is very similar, regardless of the type of wafer being processed. The wafer itself is the most important and unpredictable variable in an RTP system, as a result of its optical properties, which dominate its thermal response. This section assesses the viability of several RTP temperature measurement and control schemes, using simple heat transfer models that include the details of the thermal radiative properties of wafers. Possible temperature control schemes include pyrometry, direct contact thermocouples (TC), lamp intensity control, and the use of hot-plates. The methods used must not only fulfill the process requirements, but must also be simple to implement and calibrate, cost-effective, and robust with respect to routine fluctuations in the process and the facilities.

There are two basic approaches to control of an RTP system, closed-loop methods, which rely on the measurement of the wafer temperature for feedback to the heating control system, and open-loop methods, which rely on the repeatability of the thermal cycle experienced by the wafer. Accurate temperature measurement in an RTP system is considerably more challenging than it is in a traditional furnace. The wafer is not in thermal equilibrium with the heat source during any part of the heating cycle, and hence it is the temperature of the wafer, and not that of the furnace, which must be measured. Under open-loop control, the challenge lies in ensuring that the thermal cycle experienced by the wafer is always the same. Both closed-loop and open-loop methods of control can be affected by changes in the optical properties of the wafer, which dominate its thermal response.

11.2.7.2 Pyrometry

11.2.7.2.1 Errors and Emissivity Effects in Pyrometry

Optical pyrometry deduces the wafer's temperature from the intensity of the thermal radiation it emits at a specific wavelength [100]. There are two major problems in this approach. The first is that the pyrometer can receive stray radiation from the lamps and other system components that heat up during processing. The second difficulty is that the spectral emissivity of the wafer must be known in order to correct for deviation of the wafer's emission from blackbody behavior. Coatings on wafers can radically change wafer's spectral emissivities and cause errors in pyrometer readings [61]. The stray radiation can be minimized by various filtering methods, but an unknown target emissivity remains a problem. The error introduced when the spectral emissivity at the pyrometer wavelength, λ_p, is incorrectly set at ε_a instead of the real value, ε_r, can be calculated from the equation

$$\frac{1}{T_a} = \frac{1}{T_r} + \frac{\lambda_p}{c_2} \ln\left(\frac{\varepsilon_a}{\varepsilon_r}\right), \tag{11.30}$$

where T_r is the real temperature in K, T_a is the measured temperature in K, and c_2 is Planck's second radiation constant, which has a value of 14,388 μm K [101]. For small emissivity errors, $\Delta\varepsilon$, a simple expression can be used to assess the magnitude of the temperature error,

TABLE 11.9 Temperature Error in °C Resulting from 1% Emissivity Error for a Selection of RTP Pyrometer Wavelengths

Temperature (°C)	0.95 μm	2.4 μm	2.7 μm	3.3 μm	4.5 μm
100	0.09	0.23	0.26	0.32	0.43
200	0.15	0.37	0.42	0.51	0.70
300	0.22	0.54	0.61	0.75	1.02
400	0.30	0.75	0.84	1.03	1.41
500	0.39	0.99	1.11	1.36	1.86
600	0.50	1.26	1.42	1.74	2.37
700	0.62	1.56	1.76	2.16	2.94
800	0.76	1.91	2.15	2.62	3.57
900	0.90	2.28	2.56	3.13	4.27
1000	1.06	2.68	3.02	3.69	5.02
1100	1.24	3.12	3.51	4.29	5.84
1200	1.42	3.59	4.04	4.94	6.72

$$\Delta T \cong \frac{\lambda_p}{c_2} T^2 \frac{\Delta\varepsilon}{\varepsilon}, \tag{11.31}$$

where T is the nominal absolute temperature and ε is the nominal wafer emissivity. This expression shows that the error rises linearly with the fractional error in emissivity and with the pyrometer wavelength and that it rises with the square of the absolute temperature. It is worth noting the sign of the errors. When performing a measurement using a pyrometer, an overestimate of the wafer emissivity leads to an underestimate of the temperature. Under closed-loop pyrometer control, an overestimate of the wafer emissivity will cause the wafer to be processed at a higher temperature than desired.

Table 11.9 illustrates the impact of emissivity errors on temperature measurement at various temperatures and wavelengths. Table 11.10 shows the fractional emissivity error that is acceptable to achieve a temperature error less than 1°C for various pyrometer wavelengths and process temperatures.

Figure 11.28a illustrates the errors which would arise from the use of a pyrometer which is set to assume that the wafer emissivity is that of plain silicon, for wafers coated with oxide films of varying thickness. Figure 11.28b presents analogous results for a double layer structure in which the polysilicon layer thickness varies in a structure in which the polysilicon layer is on top of 0.2 μm of oxide. The large magnitudes of these errors illustrate the need for sophisticated approaches to reduce the impact of wafer coatings on temperature measurement in RTP.

TABLE 11.10 Emissivity Error Permissible for 1°C Temperature Error at Various Process Temperatures, for a Selection of RTP Pyrometer Wavelengths

Temperature (°C)	0.95 μm	2.4 μm	2.7 μm	3.3 μm	4.5 μm
100	12	4.4	3.9	3.2	2.3
200	7.0	2.7	2.4	2.0	1.4
300	4.7	1.8	1.6	1.3	0.98
400	3.4	1.3	1.2	0.97	0.71
500	2.6	1.0	0.89	0.73	0.54
600	2.0	0.79	0.70	0.57	0.42
700	1.6	0.63	0.56	0.46	0.34
800	1.3	0.52	0.46	0.38	0.28
900	1.1	0.44	0.39	0.32	0.23
1000	0.94	0.37	0.33	0.27	0.20
1100	0.81	0.32	0.28	0.23	0.17
1200	0.70	0.28	0.25	0.20	0.15

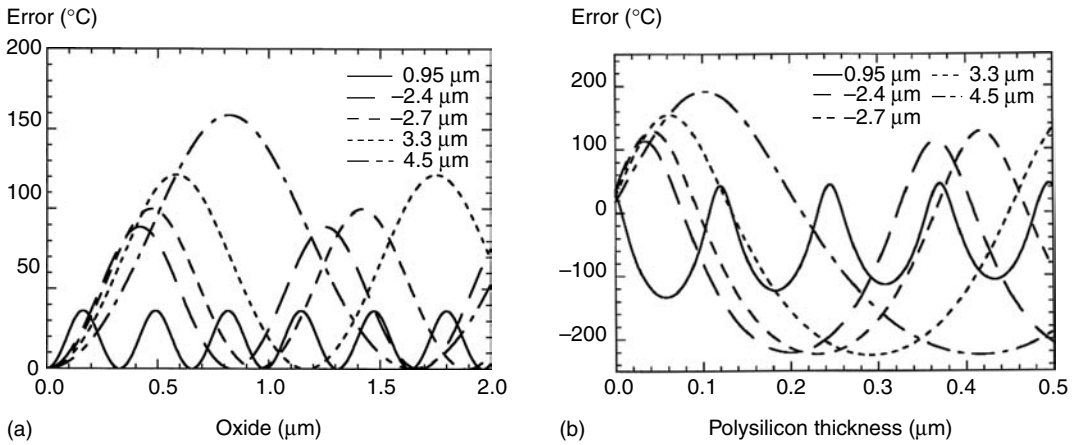


FIGURE 11.28 The impact of coating thickness on temperature measurement errors in pyrometry. The error is positive when the pyrometer overestimates the wafer temperature. Predictions are shown for several common pyrometer wavelengths. The behaviour is shown for (a) silicon dioxide films and (b) polysilicon films on top of 0.2 μm of silicon dioxide.

11.2.7.2.2 Wavelength Choice

Although the analysis above indicates that emissivity errors can be minimized by using the shortest wavelength possible, there are several other factors that influence the wavelength choice. The first aspect comes from the strength of the thermally emitted radiation. Figure 11.29 compares the temperature dependence of the Planck's function at a number of common wavelengths used for pyrometry in RTP systems. The strength of the radiation influences the minimum temperature that can be detected and also

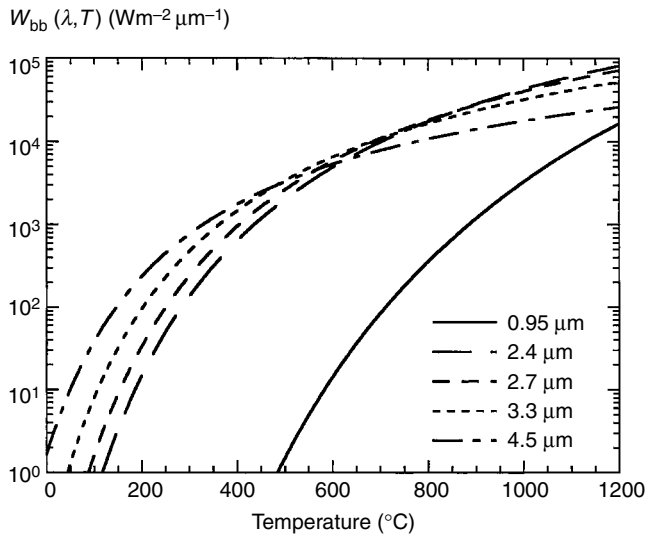


FIGURE 11.29 The temperature dependence of the spectral radiant exitance, $W_{bb}(\lambda,T)$, of a blackbody at various common pyrometer wavelengths.

the relative importance of any stray radiation that enters the pyrometer. Usually, some kind of filtering scheme is used to reduce the magnitude of the stray radiation relative to the wafer's emission. These methods can rely on absorption properties of the window between the lamps and the wafer or on mechanical light baffling schemes [7,63,102]. The choice of wavelength always ends up being a trade-off between various aspects of temperature range, stray-light elimination, emissivity-independence requirements, and signal detector capability. Another factor, which can be relevant, is the fact that lightly doped silicon wafers, which are not metallized, are usually partially transparent at temperatures below ~600°C. In real manufacturing processes, wafers are often coated with opaque films and substrates are frequently heavily doped, so this factor may not be relevant in many cases.

11.2.7.2.3 Pyrometer Calibration

Since a pyrometer relies on detecting the thermal radiation emitted at a specific waveband, a calibration procedure has to be performed to link the signal strength to the wafer temperature. The signal depends on a number of factors apart from the emissivity and temperature of the wafer, including the optical design and the way in which the pyrometer views the wafer, which is affected by the chamber geometry. For this reason, pyrometers are usually calibrated while they are on the system, using a special wafer which has thermocouples embedded in it [103–106]. This procedure takes account of all the "optics" of the system, and eliminates the need to make any assumptions about how the pyrometer views the wafer. Another calibration technique relies on the use of a signal-source to produce a fixed intensity of radiation at the pyrometer wavelength that can be linked back to the wafer temperature [107]. Other approaches have also been tried, including attempts to use melting-point temperature standards [108]. It is also possible to calibrate the pyrometer by using a monitor wafer which is known to produce a given process result when the wafer is heated through a pre-established recipe. However, this approach also has limitations that will be discussed further, below. Table 11.11 compares the advantages of different techniques of calibration.

11.2.7.2.4 Methods of Reducing Wafer Emissivity Effects in RTP Temperature Measurement

For many years, the problem of the temperature measurement errors introduced by variation in the emissivities of wafers held back the widespread use of RTP [109]. A great deal of effort was devoted to understanding the nature of the problem and attempting various solutions. Table 11.12 summarizes the relative merits of various approaches to address this problem [6,7,61,62,110–118]. Over the years, several

TABLE 11.11 Comparison of Calibration Methods

Method	Principle	Advantages	Disadvantages
Thermocouple (TC) wafer (thermocouple junction embedded within wafer)	Thermoelectric effect	Simple principle Direct link to wafer temperature Can be used over a range of temperatures	Systematic errors from imperfect thermal bond between TC and wafer Fragile Requires chamber to be opened; fine wire TC degrades over time
Radiation source	Black-body standard calibrates pyrometer signal	Mechanically durable	No direct link to wafer temperature, calibration geometry does not match process conditions Requires chamber to be opened
Monitor wafers	Known process result	Simple concept, closely linked to process requirement Direct link to wafer temperature Does not require chamber to be opened Fast technique	Different monitors may be required to calibrate at different temperatures Consumes wafers Non-thermal effects can influence monitor results

TABLE 11.12 Methods for Reducing Emissivity Errors in Pyrometry

Method	Principle	Advantages	Disadvantages	Reference
Measure wafer emissivity	Measurement of emissivity of wafer before processing	Does not require special modifications to system design	Requires a typical product wafer to be consumed.	[62]
Ex situ reflectivity (and transmissivity) measurement	Ex situ emissivity measurement using Kirchhoff's law	Does not require special modifications to system design	Highly temperature-dependent optical properties of wafers can make room temperature measurement of little value (Method likely to work better on opaque substrates) Needs to be performed on all wafers processed	[110]
Reflective cavity	Reflective enhancement of emissivity	A simple technique which greatly reduces emissivity effects	Chamber design constraint Reduces spatial resolution on highly reflecting wafers	[7,8,61,101]
In situ reflectivity (and transmissivity) measurement	In situ emissivity measurement using Kirchhoff's law	Allows emissivity corrections to be made at process temperature	Added system complexity Surface roughness effects can impact result	[6,111,112]
Ripple pyrometry	In situ emissivity measurement using modulation of heating lamps	Allows emissivity corrections to be made at process temperature	Added system complexity More difficult to apply the technique at low temperatures because of high ratio of lamp light to wafer emission signal	[113–116]
Dual effective emissivity sensors	In situ emissivity and temperature measurements deduced from output of two pyrometers with different optical configurations viewing a rotating wafer at one radius	Allows emissivity corrections to be made at process temperature	Added system complexity Surface roughness effects can impact result	[7]
Multi-wavelength pyrometry	Fit model for wafer emissivity to emission spectrum at selected wavelengths	Potentially allows more information to be used to improve temperature estimate	Adds system complexity Requires an excellent model for target spectral emissivity. Only likely to be effective when combined with in situ optical measurements	[117]
Wafer backside coating removal	Produce target with known emissivity on all wafers	Simple idea, can handle all wafers Greatly simplifies temperature measurement problem	Can add many undesirable steps to process flow	[118]

methods have demonstrated significant improvements in pyrometry for RTP. One approach involves making in situ emissivity measurements, as in the ripple pyrometry method, where an oscillating component of the heating lamp's intensity is used to measure the wafer's reflectivity and hence deduce its emissivity [113–116]. Although the idea of an in situ measurement of reflectivity seems simple, it is usually necessary to perform measurements that are independent of the wafer's surface roughness. Subtle changes in the surface texture of a wafer can result in significant temperature measurement errors and it is necessary to design the reflectivity measurement system to account for these effects [119]. If wafers are partially transparent at the pyrometer wavelength, the measurements become even more complex [111].

There are several other approaches for reducing the effect of emissivity variations. The intensity of the radiation that enters the pyrometer, is typically affected by the nature of the RTP chamber as well as by the optical properties of the wafer. This is because the chamber usually has reflecting walls, and some of the radiation emitted by the wafer is reflected between the wafer and the chamber wall several times before it enters the pyrometer. This phenomenon increases the apparent emissivity of the wafer, leading to the concept of an effective emissivity, ε_{eff}. A simple analysis of multiple reflections between plane-parallel surfaces shows that for an opaque wafer,

$$\varepsilon_{\text{eff}} = \frac{\varepsilon_{\text{r}}}{1 - R_{\text{c}}(1 - \varepsilon_{\text{r}})}, \tag{11.32}$$

where R_{c} is the reflectivity of the chamber wall and ε_{r} is the emissivity of the wafer [61]. Equation 11.32 shows that as the reflectivity of the chamber wall rises toward one, the effective emissivity tends to toward one, and the value of the wafer's free-space emissivity becomes progressively less important [8]. Figure 11.30 illustrates the effect of various chamber reflectivities on the effective emissivity. Figure 11.31 shows the effect of a highly reflecting chamber on the temperature measurement error using pyrometers with wavelengths of 0.95 and 2.7 μm for the coatings considered in Figure 11.28. Although a highly reflecting chamber wall reduces the impact of variations of wafer emissivity on temperature measurement, this calculation shows that significant errors can still arise if the emissivity of the wafer is low at the pyrometer wavelength.

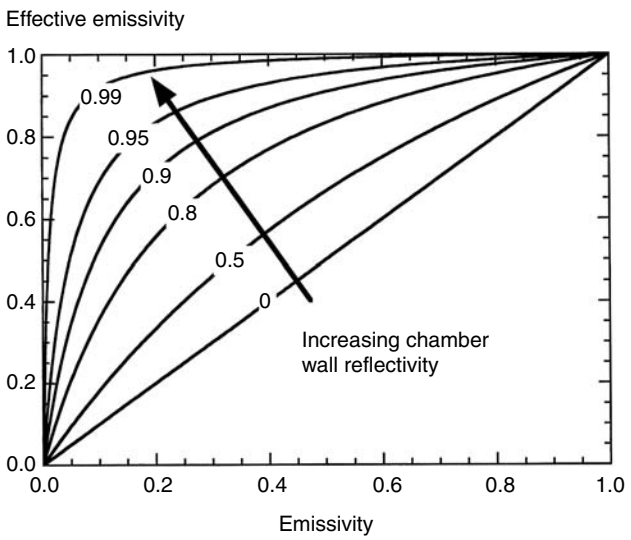


FIGURE 11.30 The relationship between wafer spectral emissivity and the effective emissivity of the wafer, as perceived by a pyrometer observing the wafer through an aperture in a reflecting surface. As the chamber reflectivity (shown by the numbers on the curves) increases, the effective emissivity approaches unity where it reaches a "virtual blackbody" condition.

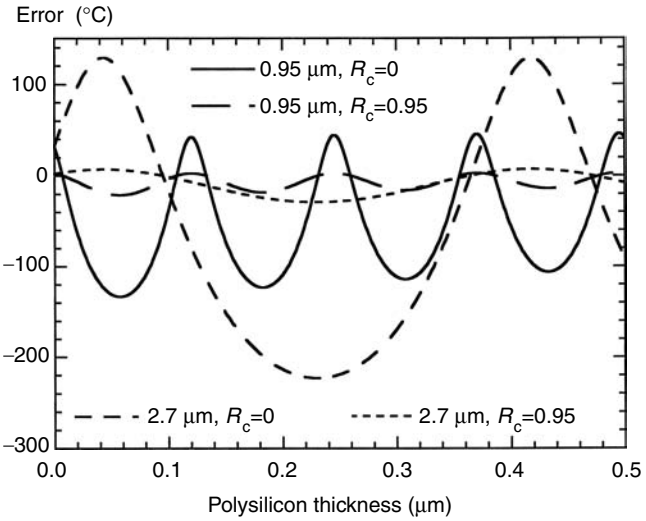


FIGURE 11.31 The effect of a reflecting chamber on the pyrometer error for polysilicon films on top of 0.2 μm of silicon dioxide. The predictions are given for two wavelengths, 0.95 and 2.7 μm, for cases where the chamber wall reflectivity, R_c, is 0 or 0.95. The calculations demonstrate how a highly reflecting chamber can greatly reduce the error.

Part of the difficulty with pyrometric approaches arises because the emissivity of some wafer backsides can change by large amounts, even for cases where the nominal coating structures are kept constant [101]. This can mean that process repeatability is degraded, even if there is no deliberate change in backside film structures. Figure 11.32a shows the temperature error introduced by a 10% change in the

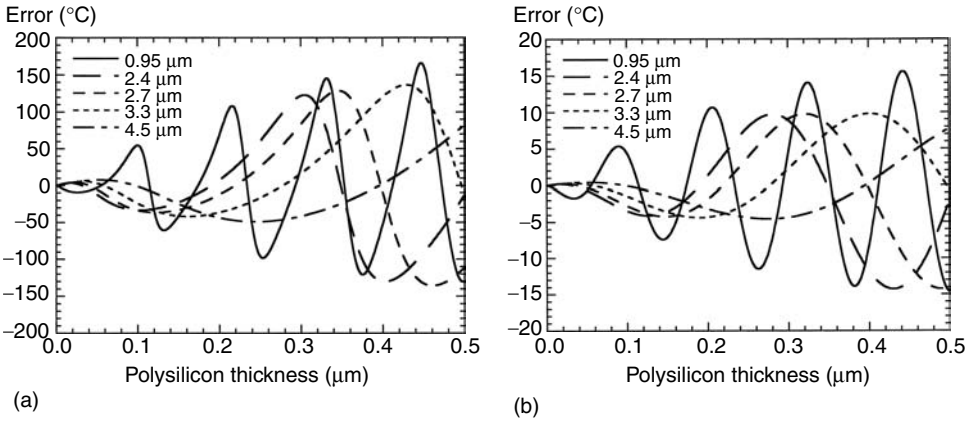


FIGURE 11.32 The impact of film thickness changes on pyrometer errors for a wafer at 1000°C. The calculations show the measurement error that would arise if a pyrometer emissivity setting was calibrated using a wafer with a given thickness of polysilicon over a 0.2-μm oxide layer, and then the same calibration was used for a wafer where the polysilicon film was 10% thicker. The calculations illustrate the strong sensitivity of pyrometer temperature control to unintentional changes in wafer emissivity, which could arise from small fluctuations in the processes that create films on the backs of wafers. The behavior is illustrated for the case where the chamber wall has a reflectivity of (a) 0 and (b) 0.95. The reflecting chamber reduces the errors.

polysilicon film thickness, when the polysilicon is on top of 0.2 μm of oxide. Figure 11.32b shows that a reflecting chamber greatly reduces the error. Nevertheless, repeatability problems of $\sim \pm 5°C$ can very easily arise in schemes where there is no in situ emissivity correction, even if there is an attempt to calibrate the emissivity.

11.2.7.3 Advanced Non-Contact Wafer Temperature Sensors

Efforts have also been made to apply various non-pyrometric temperature measurement techniques, some of which are summarized in Table 11.13 [120–130]. Most of these approaches are still in the experimental stage.

11.2.7.4 Thermocouples

The TC is one of the most widely used temperature sensors, and it can provide convenient and accurate temperature readings in a wide variety of applications. The only direct method for using a TC to establish the temperature of a wafer in an RTP system is to physically embed the TC junction within the wafer [58,103,104]. Although this technique is impractical for temperature measurement on production wafers, it is widely used for calibration of temperature sensors, setting up process recipes and troubleshooting. Wafers can also be created with a number of TCs at various positions to study the dynamic temperature non-uniformity during RTP cycles [105].

The creation of TC-wafers is not a trivial task, because the TC reports the temperature at its hot junction, which can differ from the wafer temperature if an inappropriate method is used to attach the TC to the wafer. Typically, a TC with very fine diameter leads is embedded in a small cavity in the wafer using a high-temperature cement to keep it in place. Systematic offsets between the temperature of the hot-junction and the wafer can arise because of differences between the cement's thermal properties and those of the wafer, and also because of heat loss through the TC leads. Nevertheless, a recent evaluation of TC-wafer errors suggests that they are capable of $\sim 3°C$ accuracy at 1000°C [104]. Research into more advanced techniques of TC-wafer thermometry continues, including the idea of using thin-film TCs fabricated on the wafer [106].

The most frequently used TCs are the K-type (chromel/alumel) and the R-type (platinum/platinum–rhodium). Table 11.14 contrasts the relative merits of these two types. Many issues must be considered in order to perform accurate and repeatable temperature measurements using TC-wafers, some of which are general issues in the use of TCs, and some of which are specific to the use of TC-wafers in RTP systems. Table 11.15 summarizes the main error sources and some methods for minimizing their impact. Since the sensors used for process control are often calibrated using TC-wafers, proper TC-wafer procedures can greatly improve process repeatability across time and across tools.

Although TC methods cannot be applied directly to process control, it is possible to place a sheathed TC into contact or close proximity to the wafer, and then to control the temperature of this TC. This approach is discussed in more detail below.

11.2.7.5 Open-Loop Lamp Intensity Control (OLIC)

Because of the difficulties associated with pyrometry, a number of alternative temperature control methods have evolved to provide simple, repeatable, low-cost temperature control which is robust with respect to variations in wafer properties. In these approaches, the temperature of the wafer is not controlled directly, but the target is to create highly reproducible thermal cycles.

In OLIC, the wafer is processed by programming the RTP system's lamps to run through a predetermined intensity cycle, which heats the wafer in a manner that gives the desired process results [64]. One method for establishing an OLIC recipe is to run a representative wafer through a closed-loop pyrometer or TC-controlled temperature cycle, while recording the lamp power cycle, and then to "play-back" the lamp intensity cycle when processing other wafers, whose optical properties may vary. This method allows one to create a fairly complex recipe quite easily. Tight temperature control in OLIC requires good repeatability in the heat-transfer conditions, especially the lamp power and other facilities, including the forced air-cooling flow rates and process gas flow rates [64,132,133]. One problem arises

TABLE 11.13 Non-Pyrometric Methods for Measuring the Temperatures of Silicon Wafers

Method	Principles	Issues	Reference
Grating	Measurement of thermal expansion of silicon by observing movement of diffracted laser beam illuminating grating on wafer	Grating has to be fabricated on each wafer Beam deflection angles are very small	[120]
Wafer extension	Measurement of thermal expansion of silicon by observing change in wafer diameter	Only gives integrated temperature profile across wafer Vulnerable to errors from wafer deflection and warping	[121]
Ultrasonic—contact	Measurement of speed of sound using pin supports to excite and detect ultrasonic wave propagation in wafer. Tomographic approach permits crude uniformity measurement	Non-tomographic approach only gives average temperature reading Requires special pin support design Vibration can affect results	[122]
Ultrasonic—non-contact	Measurement of speed of sound using laser pulse to excite wave and interferometer to detect wave	Trade-off between excitation pulse power, spatial resolution and temperature resolution	[123]
Absorption edge sensing	Measurement of shift of wafer absorption edge by transmission or reflection spectrum	Wafer must be transparent and of known doping concentration Typically limited to $<800°C$ by strong absorption in silicon	[24,124]
IR transmission	Measurement of wafer transmission at selected wavelengths	Wafer must be transparent and of known doping concentration Typically limited to $<800°C$ by strong absorption in silicon	[125,126]
Reflectivity	In situ measurement of wafer reflectivity	Requires a surface with known optical properties Small effect, requires special instrumentation and careful optical alignment Potentially sensitive to surface roughness and surface stability	[127]
Ellipsometry	In situ measurement of optical constants	Requires a surface with known optical properties Small effect, needs special instrumentation and careful optical alignment Potentially sensitive to surface roughness and surface stability	[128]
Interference effect	Determination of "optical thickness" of a wafer. Linked to temperature by known T dependence of optical constants and expansion effect	Best results are on double-side polished wafers Typically limited to $<800°C$ by strong absorption in silicon	[129]
Raman spectroscopy	In situ measurement of Raman spectrum, linked to temperature by basic solid-state physics	Formidable instrumentation and interpretation problems Effects of doping and stress can lead to complex phenomena which are difficult to interpret	[130,131]

T, Temperature.

TABLE 11.14 Comparison of Thermocouple Types for Thermocouple-Instrumented Wafers

Thermocouple	Composition	Advantages	Disadvantages
K	Chromel (Cr/Al)–Alumel (Ni/Al)	Sensitive Low thermal conductivity (reduces heat loss from junction) Mechanically strong Low cost	Sensitive to oxygen "Special grade" wire only has $\pm 0.4\%$ accuracy Inhomogeneous short-range ordering can cause problems at 250–550°C range. Effect is reversible
R	Pt/Rh (13%)—Pt	Resists oxidation "Special grade" wire has $\pm 0.1\%$ accuracy	Mechanically weak High thermal conductivity (increases heat loss from junction) Low sensitivity Very reactive with silicon Expensive

The R-type thermocouple is in many ways similar to the S-type thermocouple which is also in use.

from the tendency of the quartz components, and especially windows or showerheads, to progressively heat up as more and more wafers are processed in the chamber [89,134,135]. Progressive changes in quartz temperature can influence both wafer temperature repeatability and uniformity. The problem can be minimized by preheating the quartzware, either by heating with the lamps alone, or more effectively, by preheating the system with dummy wafers.

Initially, it was hoped that OLIC would provide immunity to variations in coatings, however, it turns out that these coatings can change the radiative properties enough to induce large temperature shifts [64]. However, there can be a practical benefit in using OLIC because the technique allows a simple method for reducing the temperature variations caused by small changes in the thicknesses of the coatings introduced by fluctuations in the process producing these surface coatings. These variations can cause very large errors in conventional pyrometric temperature control because of their large impact on the spectral emissivity at the pyrometer wavelength. In OLIC, their effect is averaged out over the wide wavebands of thermal radiation involved in determining the wafer temperature, and their influence is much smaller [101].

The effect of coatings on OLIC can be analyzed using the simple model for lamp heating described in Section 11.2.6.4 combined with integrated radiative properties calculated using the methods of Section 11.2.5 [57,101]. Figure 11.33 shows the predicted transient response of a lightly doped wafer with no coatings when a fixed lamp intensity is applied for 60 s, as compared to one with a two-layer polysilicon-on-oxide coating on one of its surfaces. In the example shown, the total lamp intensity is 100 kWm^{-2}. The coatings alter both the transient response and the steady-state temperature. In this case, the plain wafer reaches 1011°C at the end of the 60 s heating cycle, while the wafer with the coating on reaches 973°C. The thermal response of the quartz isolation tube is also altered by the presence of different coatings on the wafers surfaces.

Figure 11.34a shows the temperatures reached under the same conditions for various thicknesses of oxide on a wafer and also for various thicknesses of polysilicon on top of a 0.2-μm thick oxide layer. The calculations were performed using a slightly simplified thermal model, which did not include the dynamic response and which fixed the quartz tube temperature at 400°C and the lamp temperature at 2500°C. The lamp intensity was chosen so that a plain silicon wafer would reach 1000°C. Predictions are shown based on the use of both integrated normal and hemispherical properties. The calculations based on hemispherical integrated properties show a slightly smaller range of temperatures, but the range is still ~60°C which illustrates the necessity of calibrating the OLIC recipe for the product wafer of interest. Figure 11.34b shows the steady-state temperature changes expected for a 10% increase in the thickness of the oxide film or a polysilicon film on top of a 0.2-μm thick oxide layer on one side of a wafer, relative to the values shown on the *x*-axis. The resulting temperature changes are all less than ~8°C, demonstrating that the OLIC method is reasonably resilient to process-related fluctuations in wafer surface coatings.

TABLE 11.15 Errors in the Use of Thermocouples (TCs)

Type of Error	Cause of Error	Method for Minimizing Error	Comments
Thermal error	Difference in temperature between TC and point being measured	Intimate thermal bond between TC and object Minimize wire diameter and thermal conductivity to reduce heatsinking effect of wire	In an RTP chamber the wafer is not in thermal equilibrium with the chamber. TCs must be physically embedded within the wafer to obtain accurate readings
Wire errors	Variations in the thermoelectric characteristics of the TC wires	Use high grade wire Calibrate the TC against another standard	Even high grade wire can have a fairly loose specification, e.g., special grade K-type wire has an error of $\sim \pm 4°C$ at 1100°C. For the fine TC-wires used in most TC-wafers, the usual specifications are often not guaranteed
Wire drift	Changes in physical characteristics of wire over time	Operate the wire in correct ambient Limit the number of thermal cycles applied to any TC	For K-type TCs, the oxygen concentration in the chamber should be minimized by thorough purging Up to $\sim$200 cycles have been demonstrated on K-type TC-wafers with $< \sim 2°C$ drift
Virtual junction	Low resistance path between TC wires at some point other than junction	Use adequate insulation Lay out wires to prevent accidental shorting	
Extension and compensating wire	Poor match between thermoelectric characteristics of wire and TC causes temperature error	Use correct grades of wire to connect TC to electronics Minimize thermal gradient across extension wire	Great care must be taken with connection schemes
Electronics errors	Voltage measurement errors	Optimize electronics	TC voltages are very small and measurements are prone to problems with interference (pick-up) and noise
	Reference junction correction	Incorporate reference junction in electronics	
	Voltage to temperature errors	Use adequate conversion algorithm to interpret voltage as temperature	
	Noise rejection	Pay careful attention to signal grounding and shielding issues	

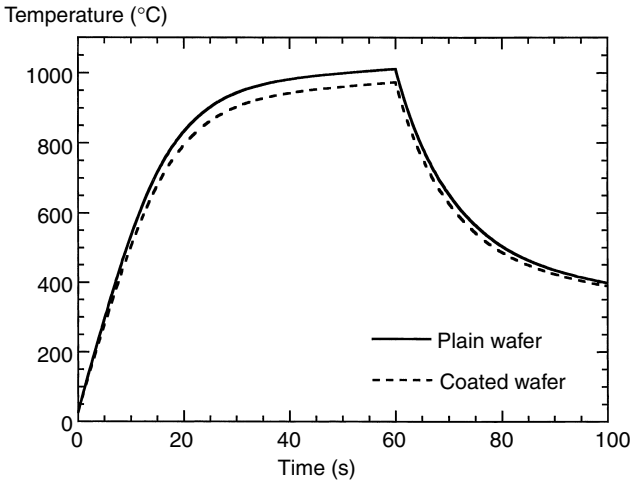


FIGURE 11.33 A prediction of the effect of a coating on the thermal response of a wafer exposed to 60 s heating with fixed-intensity lamp radiation with a total power density of 100 kWm^{-2}. The coating is a 0.2-μm film of polysilicon on top of a 0.2-μm film of silicon dioxide.

Figure 11.35a shows a contour map of the open-loop temperatures predicted for various polysilicon-on-oxide coating structures for the case where plain silicon would reach 1000°C. In this calculation, the quartz temperatures were also allowed to vary, and the convective cooling coefficient for the quartz was set to 70 $Wm^{-2}K^{-1}$. The result shows the strong influence of polysilicon-on-oxide films on the thermal

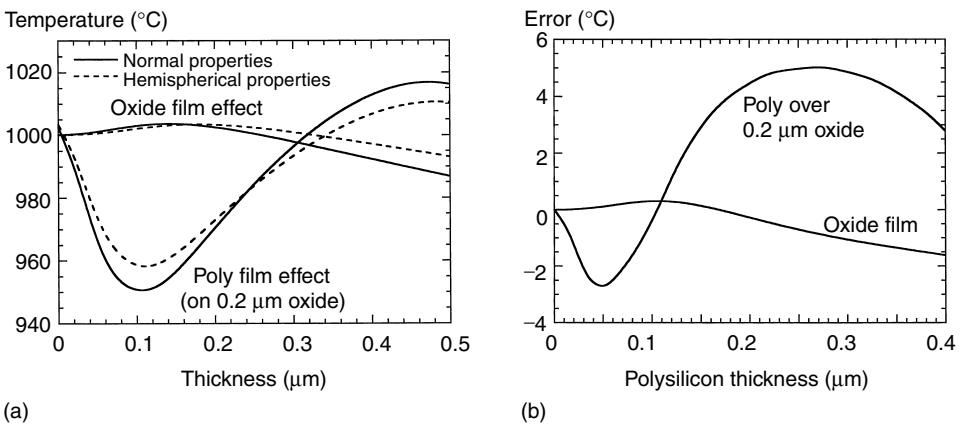


FIGURE 11.34 (a) Predictions of the effect of coating thicknesses on temperatures reached during open-loop lamp intensity control (OLIC). One pair of curves shows the effects of oxide film thickness for a wafer coated with a silicon dioxide film. The other curves show the effects of polysilicon film thickness for a wafer that has a coating with a polysilicon film over 0.2 μm of silicon dioxide. The lamp radiation intensity was chosen so that a plain silicon wafer would reach 1000°C. Calculations were performed using either normal integrated radiative properties or hemispherical integrated radiative properties. The predictions demonstrate that OLIC requires calibration. (b) The effects of film thickness changes on wafer temperatures for OLIC that would heat a plain silicon wafer to 1000°C. One curve shows the impact of a 10% increase in the film thickness for the oxide film on an oxide-coated wafer, and the other shows the effect for a 10% increase in the polysilicon film thickness for a wafer with polysilicon over 0.2 μm of oxide. The wafer temperature under OLIC is reasonably resilient to small fluctuations in coating structures.

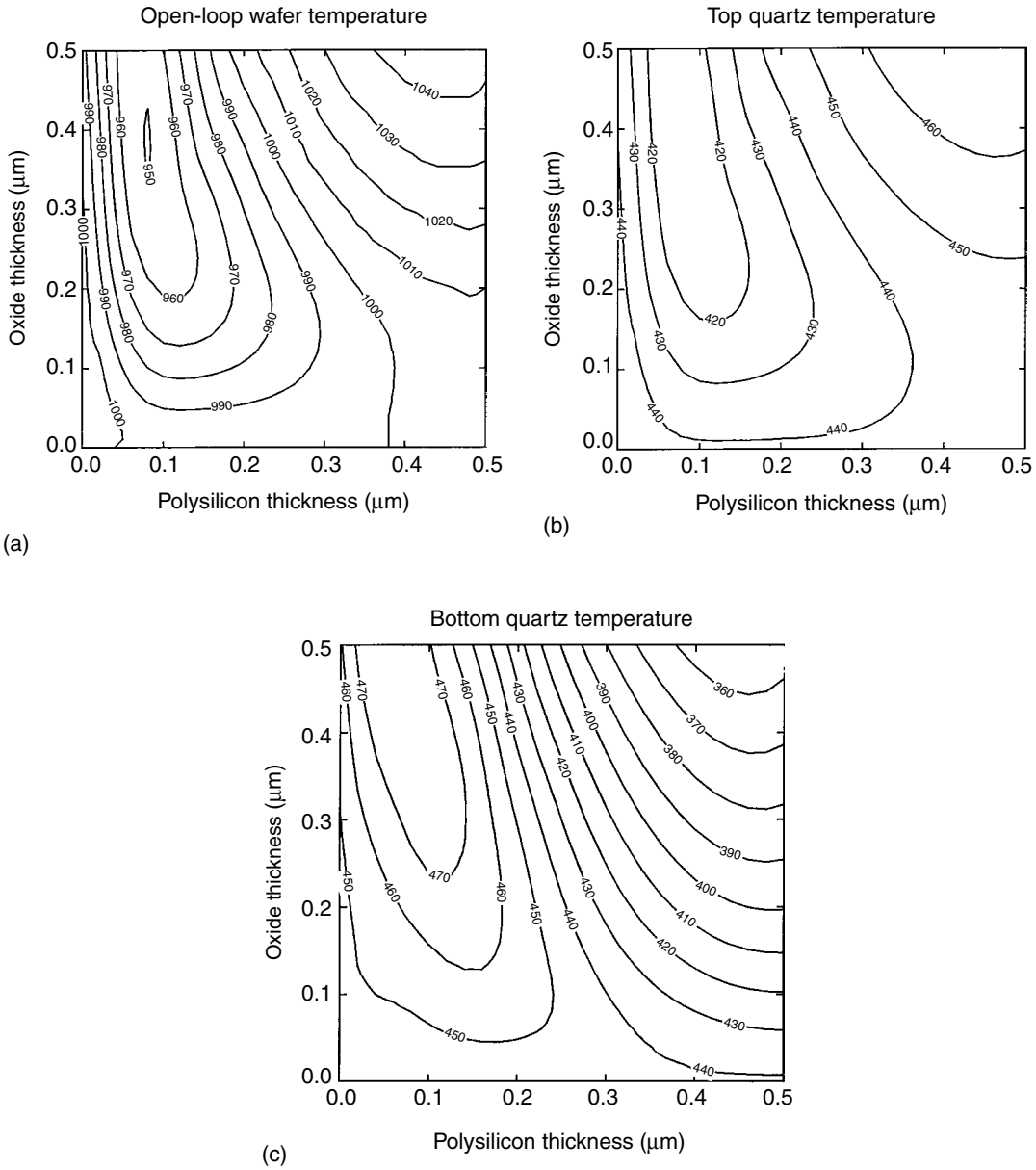


FIGURE 11.35 (a) Contour map showing predictions of the wafer temperatures that would be reached on wafers with coatings consisting of polysilicon on top of silicon dioxide, during OLIC with lamp power intensity set to obtain 1000°C on plain silicon. Changes in the film thicknesses can result in large fluctuations in the wafer temperature. (b) Predictions of the temperature of the quartz plate above the wafer follow a pattern similar to the trend for the wafer temperature. (c) Predictions of the temperature of the quartz plate below the wafer show that its temperature is more strongly affected by the thermal radiative properties of the wafer than by the wafer temperature itself.

radiative properties and consequent heating behavior. There is ~100°C change in wafer temperature for the range of oxide and polysilicon thicknesses covered here. The coating also affects the quartz heating. Figure 11.35b shows that the temperature of the top quartz, which faces bare silicon, varies from ~415 to 465°C as the wafer temperature varies. Figure 11.35c shows the behavior for the lower quartz plate, which

faces the coated silicon surface. Here, the trend is quite different, because the coating changes the long-wavelength-band emissivity of the wafer, which determines how efficiently the wafer heats the quartz. The temperature varies from 360 to 470°C, and the quartz temperature is high when the wafer temperature is low and vice versa.

11.2.7.6 Hot-Plates

Rapid thermal processing systems that use a hot-plate as the heating source essentially operate in an open-loop heating mode, where the temperature of the hot-plate is controlled rather than that of the wafer. The wafer is loaded onto pins which lower it to a position close to the surface of the hot-plate as shown in Figure 11.1d [10,11].

Figure 11.36 illustrates the heat transfer process when a wafer is heated using a hot-plate [57]. There is no lamp heating, and the wafer is heated by a combination of thermal radiation and conduction through the process gas. The distance between the wafer and the hot-plate determines the ratio of the energy transfer by these two mechanisms. An analysis can be performed using a simple model, much like that used in the case of the assessment of OLIC above [57]. For the calculations shown here, it was assumed that the hot-plate is a spectrally gray surface with an emissivity of 0.8 which is kept at fixed temperature during processing. Figure 11.37 shows predictions of the temperature cycles for lightly and heavily doped wafers which are at room temperature when they are loaded onto a hot-plate that is kept at 1100°C. Predictions are shown for cases where the wafer is 1 or 2 mm from the hot-plate in a nitrogen ambient, and for a vacuum ambient, where the gap between the wafer and the hot-plate is irrelevant, because there is no heat conduction. The heating cycles for the lightly and heavily doped wafers are different because of their very different optical properties. Lightly doped wafers in a vacuum ambient only reach their steady-state temperature after $\sim$200 s of heating. In nitrogen, heat can be transferred by conduction, which should reduce the influence of the wafer's optical properties on the thermal cycles. Figure 11.37 shows that the heating rates are considerably higher in a nitrogen ambient than in vacuum.

Figure 11.38a presents the predicted steady-state wafer temperature reached for a heavily doped plain silicon wafer when it is loaded onto a hot-plate which is at various distances away from the wafer in a gas ambient with either a high (0.5 $Wm^{-1} K^{-1}$) or a low (0.05 $Wm^{-1} K^{-1}$) thermal conductivity. The latter case is typical of nitrogen gas [57]. The figure also shows the temperature differences in a vacuum. Since the model does not include the convective heat transfer, it is expected to become inaccurate if the gap between the wafer and the hot-plate exceeds a few millimeters. It is clear, however, that the wafer

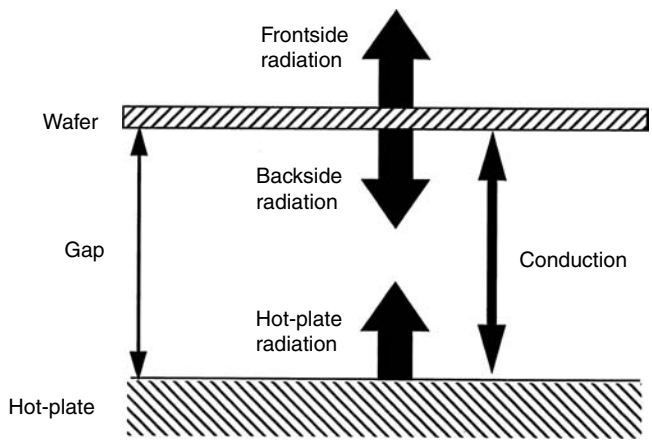


FIGURE 11.36 A simple model for heat transfer from a hot-plate to a wafer.

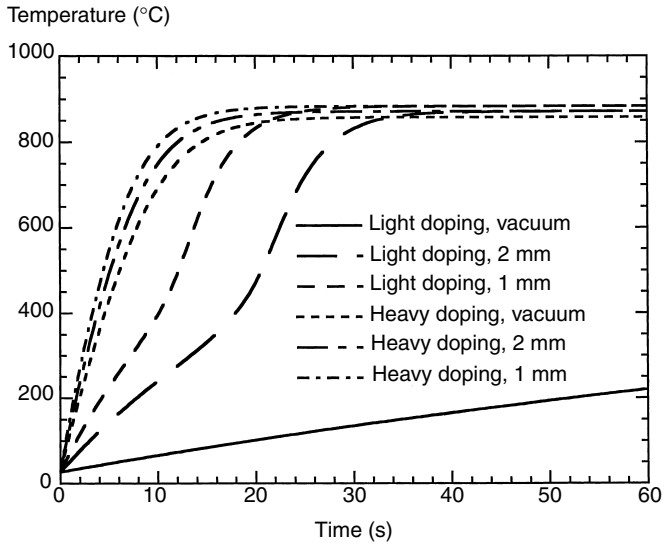


FIGURE 11.37 Predictions of wafer heating transients for wafers loaded onto a hot-plate that is kept at 1100°C. The results are calculated for wafers with light or heavy doping, for either 1 or 2 mm gaps between the wafer and the hot-plate in a nitrogen ambient, and for vacuum conditions. Conduction through the gas has a strong impact on the thermal response.

temperature will be considerably lower than that of the hot-plate, with a difference of ~240°C for operation in a vacuum. In commercial RTP systems based on the hot-plate approach, attempts are made to reduce the heat loss from the front of the wafer to minimize the tendency for a large temperature difference to arise between the wafer and the hot-plate [11]. The use of small wafer/hot-plate gaps and highly conductive gases such as helium can also help reduce the temperature difference. Figure 11.38a

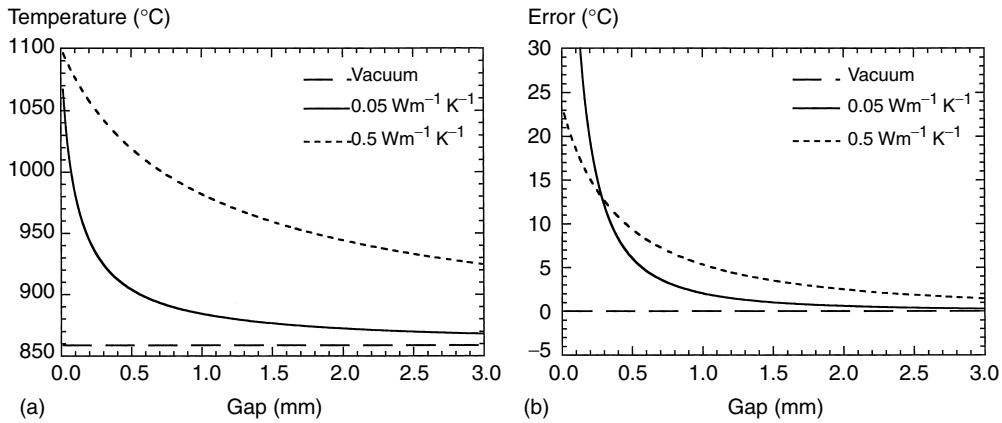


FIGURE 11.38 (a) The effect of the separation between a wafer and a hot-plate on the wafer temperature. The hot-plate is at 1100°C, and predictions are presented for gas ambients with thermal conductivities of 0.05 or 0.5 $Wm^{-1}K^{-1}$ and for the case of a vacuum ambient. In the vacuum ambient there is no thermal conduction, and the wafer/hot-plate separation has no effect on the wafer temperature. (b) The effect of a 0.1-mm change in the wafer/hot-plate gap on the predicted wafer temperature. The size of the gap has a very strong effect on the wafer temperature for gaps <0.5 mm.

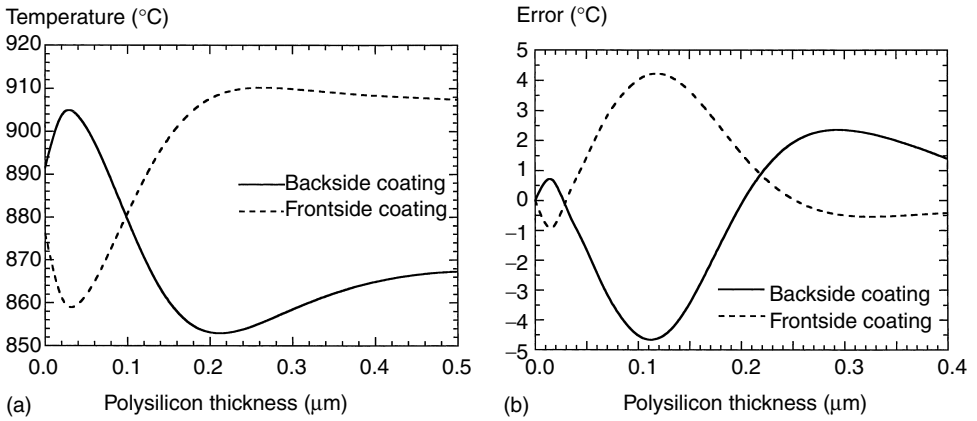


FIGURE 11.39 (a) The effect of coating film thickness on the wafer temperature reached during heating by a hot-plate. The calculations show the effects of polysilicon film thickness for a wafer that has a coating with a polysilicon film over 0.2 μm of silicon dioxide. The behavior is shown for the cases where the coating is on either the front or the back of the wafer. The hot-plate is at 1100°C, and the wafer is 1 mm above it, in a nitrogen ambient. The calculations show that this approach is sensitive to the nature of the wafer coatings. (b) The effect of film thickness changes on the wafer temperature during hot-plate heating. The calculations demonstrate the impact of a 10% increase in the film thickness of a polysilicon film for a wafer coated with polysilicon on top of 0.2 μm of oxide. The results show that the wafer temperature is reasonably stable with respect to small fluctuations in coating structures.

shows that a gas ambient greatly reduces the difference as the wafer is brought close to the hot-plate, but if the wafer is very close to the hot-plate, the behavior becomes strongly dependent on the size of the gap. This could cause practical problems related to the mechanical tolerance on this gap, wafer alignment to the hot-plate and the finite warp of either the wafer or the hot-plate. Figure 11.38b shows the effects of a 0.1-mm change in the spacing for various wafer/hot-plate separations. The results suggest that control of this gap becomes critical to maintaining temperature uniformity once the gap is less than 1 mm, and calculations also suggest that large dynamic temperature difference could arise across a wafer as it is loaded onto a hot-plate [57].

It is interesting to assess whether this approach can eliminate the influence of wafer emissivity effects on the temperature cycle. Figure 11.39a shows predictions of the temperatures reached when coated wafers are loaded onto a hot-plate maintained at 1100°C in a nitrogen ambient with a 1-mm gap between the wafer and the hot-plate. The wafer's backsides or frontsides are coated with various thicknesses of polysilicon on top of a 0.2-μm thick oxide layer. The results demonstrate a spread in temperatures of $\sim$60°C, and show that it is necessary to adjust the process recipe for the product wafer of interest. Figure 11.39b shows the steady-state temperature changes, which would result from a 10% increase in the thickness of a polysilicon film on top of a 0.2-μm thick oxide layer on either side of a wafer, relative to the values shown on the *x*-axis. The resulting temperature changes are all in the range $\sim \pm 5$°C, and as with the OLIC method, this approach can be expected to be reasonably stable with respect to process-related fluctuations in wafer surface coatings.

11.2.7.7 Radiation Shields

Rapid thermal processing using lamps can also be performed with the configuration shown in Figure 11.40, where an opaque structure under the wafer, called a radiation shield, masks the wafer backside from the lower lamps, and provides a target whose temperature can easily be monitored by a pyrometer or a TC [101,136,137]. The wafer can also be placed between two shields, so that it does not see the lamps at all. The latter method is often used for processing GaAs wafers, partly to provide

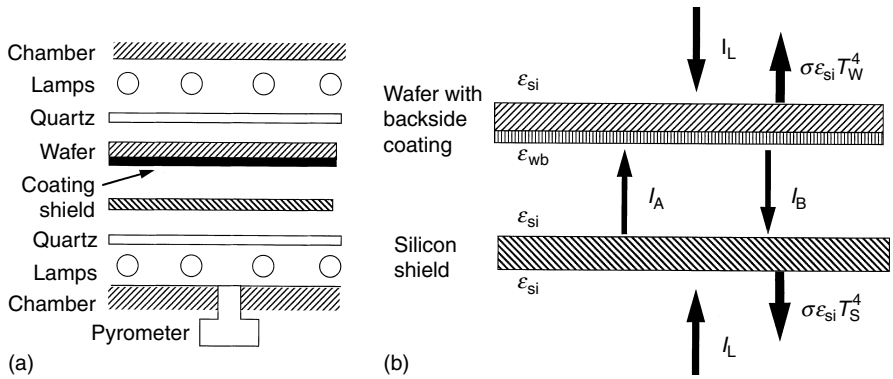


FIGURE 11.40 (a) A radiation shield in a double-side heating RTP configuration provides a simple method for making temperature control independent of wafer backside emissivities. (b) The operation of the shield can be explained by a simple heat-transfer model. If the bottom surface of the shield is silicon, and the top surface of the wafer is silicon, and the lamp power is incident equally from above and below the wafer, the symmetry of the heat transfer problem requires the steady-state wafer temperature to equal that of the shield, regardless of the wafer backside coating.

a semi-enclosed environment where a local arsenic overpressure can be generated, and partly to increase lamp coupling and simplify temperature measurement [138]. Silicon processing may also benefit from this approach, especially from the point of view of temperature control.

The effect of a controlled heating cycle in which the shield temperature is ramped up to a process temperature at a fixed rate was simulated by predicting the lamp intensity cycle necessary to perform this ramp. Figure 11.41 shows predicted wafer responses when the radiation shield is spectrally gray with an emissivity of 0.8, and it has the physical properties of silicon. The shield is heated under closed-loop control from room temperature to either 700 or 1000°C at a ramp rate of 50°C/s. The shield and the wafers are 725 μm thick, and the shield is 6 mm below the wafer in a nitrogen ambient. The gap between the wafer and the quartz was set equal to that between the shield and the quartz, at 7 mm. The response of a heavily doped wafer is very similar to that of the shield, but the lightly doped wafers heat up slower than the shield. Because the doping influences the thermal cycle, the system has to be optimized separately for different kinds of wafers. This is a consequence of the open-loop nature of this heating arrangement, in which nothing senses the wafer's temperature, and it is controlled via an intervening shield [57].

The shield eliminates the backside sensitivity of the system by "hiding" the thermal radiative properties of the wafer backside from the heat-transfer problem. A key aspect of the method relies on the inherent symmetry between the wafer and the shield, which can be achieved double-sided heating configuration. It is easiest to understand the method in the case where the top surface of the wafer is bare silicon, and the lower surface of the shield is also bare silicon. If the power delivered by upper and lower lamp banks is equal, then the system is symmetrical and there is no net heat transfer through the structure. This makes the emissivities of the internal surfaces in the wafer/shield "sandwich" irrelevant, and the wafer backside coating does not affect the temperature. Mathematically, the argument can be summarized using the notation in Figure 11.40b, starting with the expression for the power balance on the wafer at steady-state,

$$I_L - \sigma\varepsilon_{si}T_w^4 + (I_A - I_B) = 0, \tag{11.33}$$

where I_L is the lamp power density, ε_{si} is the total emissivity of silicon, T_w is the wafer temperature and the term $(I_A - I_B)$ represents the interchange of energy between the wafer and the shield. This is given by

$$I_A - I_B = \frac{\sigma\varepsilon_{si}\varepsilon_{wb}}{1 - R_{si}R_{wb}}\left(T_s^4 - T_w^4\right), \tag{11.34}$$

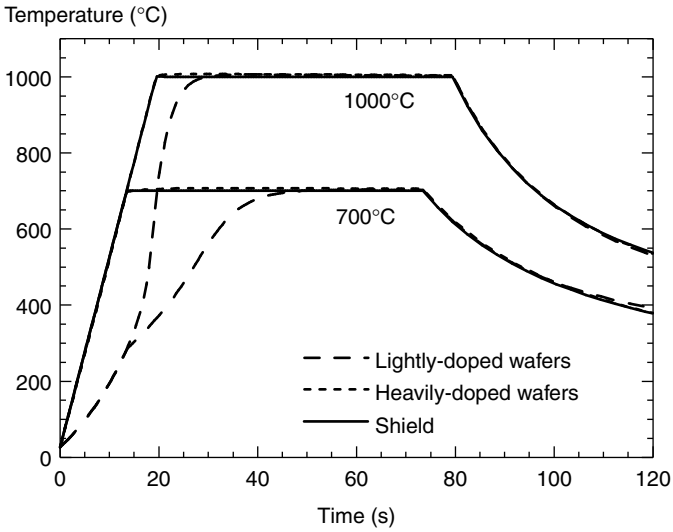


FIGURE 11.41 Calculated temperature transients for RTP with a radiation shield. Predictions are shown for an opaque shield heated under closed-loop control to either 700 or 1000°C. Predictions are shown for lightly and heavily doped wafers. The shield was assumed to have the thermal properties of silicon and to have a spectral emissivity of 0.8 at all wavelengths. The thermal response of the shield and the heavily doped wafers is very similar, but for the lightly doped wafer significant differences between the radiative properties of the shield and the wafer result in different transient thermal responses. The steady-state temperatures are not affected by the doping because the wafers are all opaque at temperatures $> \sim 650$°C.

where T_s is the shield temperature, ε_{wb} is the total emissivity of the wafer backside, and R_{si} and R_{wb} are the total reflectivities of silicon and the wafer backside, respectively. For the shield, the steady-state power balance is summarized by

$$I_L - \sigma \varepsilon_{si} T_s^4 - (I_A - I_B) = 0. \tag{11.35}$$

I_L, I_A, and I_B can be eliminated from these expressions, and rearranging one obtains

$$\sigma \left(\varepsilon_{si} + \frac{2\varepsilon_{si}\varepsilon_{wb}}{1 - R_{si}R_{wb}} \right) \left(T_s^4 - T_w^4 \right) = 0, \tag{11.36}$$

which implies that $T_s = T_w$ for all values of ε_{wb}. If the top surface of the wafer, the bottom surface of the shield, or the lamp bank ratio change, then the temperatures of the wafer and the shield cease to be equal. As a result, a simple implementation of this approach does not completely compensate for the effect of product wafers with different frontsides. However, the wafer frontside is not usually one continuous coating and its impact as a variable should be less significant than that of the backside. It is also important for the shield to maintain stable mechanical and optical properties over many thermal cycles. Other practical problems arise from the open-loop nature of the control of wafer temperature. For example, there may be drift in the temperature calibration because of quartz tube clouding or lamp aging. Significant improvements in the use of a radiation shield are possible by combining it with direct TC control, as described below.

Figure 11.42a shows the steady-state temperatures predicted for coated wafers processed in a system in which the shield considered earlier is heated to 1000°C. The wafer's backsides are coated with various

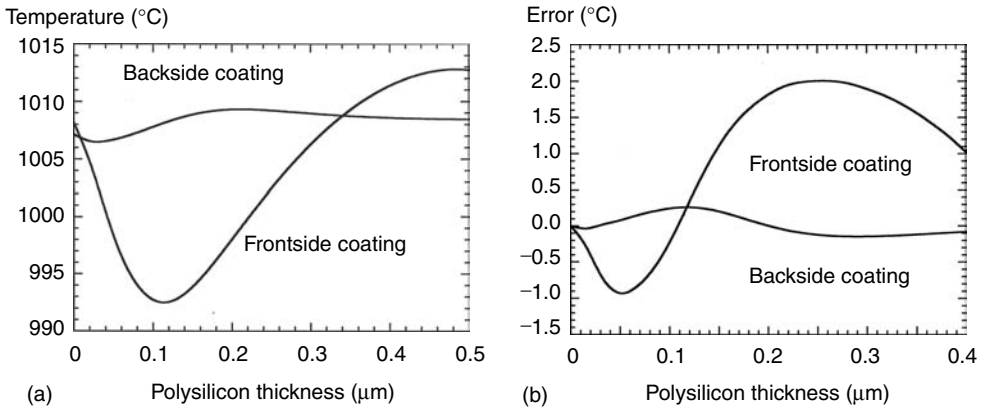


FIGURE 11.42 (a) The effect of coating film thickness on the wafer temperature reached in a double-side heating RTP system with a radiation shield. The calculations show the effects of polysilicon film thickness for a wafer that has a coating with a polysilicon film over 0.2 μm of silicon dioxide. The behavior is shown for the cases where the coating is on either the front or the back of the wafer. The radiation shield is assumed to be at 1000°C and to be opaque and spectrally grey, with an emissivity of 0.8. The calculations show that the approach virtually eliminates the effect of the backside coatings, but some effect remains from variations in the frontside coating. (b) The effect of film thickness changes on the wafer temperature. The calculations demonstrate the impact of a 10% increase in the film thickness of a polysilicon film for a wafer coated with polysilicon on top of 0.2 μm of oxide. The results show that changes in the wafer coatings have very little impact on the process temperatures.

thicknesses of polysilicon on top of a 0.2-μm thick oxide layer. The calculations were performed using a slightly simplified thermal model, which did not include the dynamics and in which the quartz tube temperature was fixed at 400°C. The results demonstrate a temperatures range of <3°C when the backside is coated, and illustrate excellent emissivity-independence. The frontside coating shifts the temperature, introducing a variation of ~ 15°C. Figure 11.42b shows that the method is still very resilient against changes in coatings, because the $+10\%$ change in the frontside polysilicon film thickness introduces $< \pm 2$°C temperature change. Experimental studies have confirmed the excellent emissivity independence which can be attained by using radiation shields [137]. The results above suggest that a radiation shield approach could be very useful in a manufacturing environment where many different types of wafer might be being processed and large changes in the backside film composition could routinely occur.

11.2.7.8 Direct Thermocouple Control

A TC inside a sheath can be used for temperature control by placing it in direct contact with the wafer [139–141]. The sheath, which is usually made of SiC or quartz, protects the wafer from contamination from the TC, yet it can be made thin enough to have a thermal response time short enough to track wafer temperature fluctuations. This kind of temperature probe receives energy from the wafer by a mixture of radiation and thermal conduction through the process gas as well as radiation from the lamps and quartz tube. The influence of the various energy sources on the temperature reading can be taken into account by a calibration procedure that compensates for the temperature offset between the probe and the wafer. Once the probe is calibrated for control on a given type of wafer, it is insensitive to the typical process-related fluctuations in the thicknesses of the films on the back of the wafer [139,140]. Furthermore, any heat which is transferred by thermal conduction through the process gas is independent of the optical properties of the wafer's back surface.

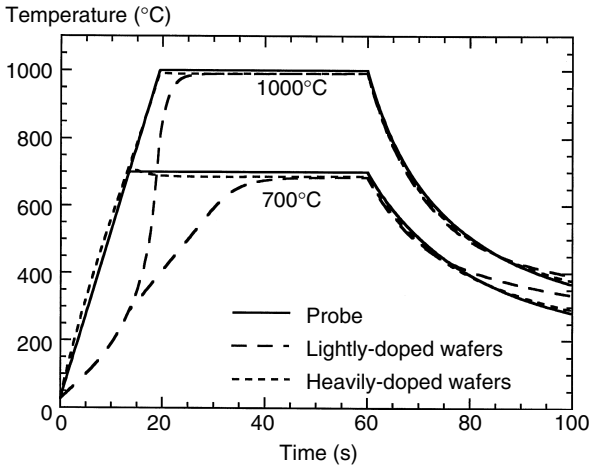


FIGURE 11.43 Calculated temperature transients for RTP under the control of a thermocouple (TC) probe. Predictions are shown for an opaque TC probe that is heated under closed-loop control to either 700 or 1000°C. The calculations are shown for both lightly and heavily doped wafers. The TC probe is assumed to have a spectral emissivity of 0.8 at all wavelengths, and its thermal mass is chosen so that the transient response of the probe and the heavily doped wafers is similar. For the lightly doped wafers, the significant differences in radiative properties between the probe and the wafers result in different transient thermal responses. However, the steady-state temperatures are not affected by doping, because the wafers are all opaque at temperatures $> \sim 650$°C.

The behavior of the probe can be analyzed using methods very similar to those discussed above. Although the geometry of the probe does not closely resemble the infinite parallel planes geometry, it turns out that excellent agreement can be obtained between theory and experiment even with this approximation. In the model used here, the effect of conduction through the process gas between the wafer and the probe is ignored, because at high temperatures thermal radiation dominates the energy exchange. Figure 11.43 shows the predicted responses of both lightly and heavily doped plain silicon wafers that are 725 μm thick, when they are heated under the control of a TC probe. The probe is assumed to be spectrally gray with an emissivity of 0.8, and its thermal mass was chosen to make its response similar to that of the wafer. The results are shown for cases where the probe temperature is ramped to either 700 or 1000°C at a rate of 50°C/s. As with the radiation shield results shown in Figure 11.41, it is evident that the behavior for lightly doped wafers is quite different to that of heavily doped material during the ramp-up.

The model can also be used to predict the impact of wafer coatings on the temperature control. Figure 11.44a shows predictions of the steady-state temperatures expected for coated wafers processed under conditions where the probe reaches a temperature of 1000°C. The wafer's backsides are coated with various thicknesses of polysilicon on top of a 0.2-μm thick oxide layer. The calculations were performed using a slightly simplified thermal model, which did not include the dynamics and fixed the quartz tube temperature at 400°C. The results show a ~ 55°C temperature range and illustrate the need to calibrate the thermal cycle. Figure 11.44a also shows that the same coating variations on the frontside of the wafer only shift the temperature through a range of ~ 10°C. Figure 11.44b shows that the method is resilient against fluctuations in coatings, because a $+10\%$ change in backside polysilicon film thickness only introduces temperature changes $\sim \pm 3$°C.

11.2.7.9 Combined Radiation Shield and Direct Contact Thermocouple

For even greater resilience against variations in wafer properties, it is possible to combine a radiation shield with a direct TC control approach. In this configuration, when the shield/wafer sandwich is heated

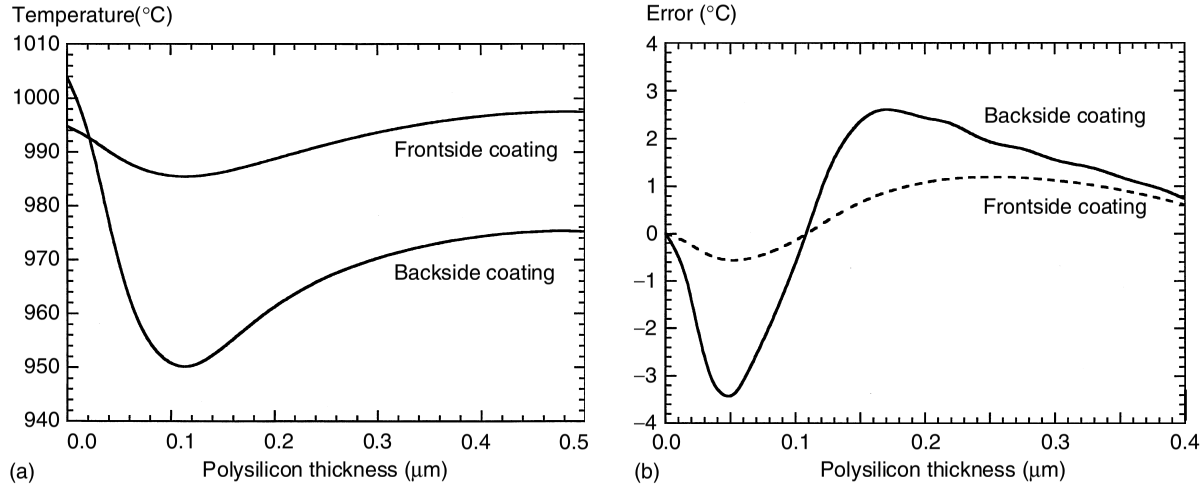


FIGURE 11.44 (a) The effect of coating film thickness on the wafer temperature reached during heating under the control of a TC probe. The calculations show the effects of polysilicon film thickness for a wafer that has a coating with a polysilicon film over 0.2 μm of silicon dioxide. The behavior is shown for the cases where the coating is on either the front or the back of the wafer. The probe is assumed to be at 1000°C and to be opaque and spectrally grey, with an emissivity of 0.8. The calculations show that the wafer temperature is affected by the coating, and that this approach to temperature control requires calibration. (b) The effect of film thickness changes on the wafer temperature. The calculations demonstrate the impact of a 10% increase in the film thickness of a polysilicon film for a wafer coated with polysilicon on top of 0.2 μm of oxide. The results show that the TC probe approach is reasonably resilient to changes in the wafer coatings.

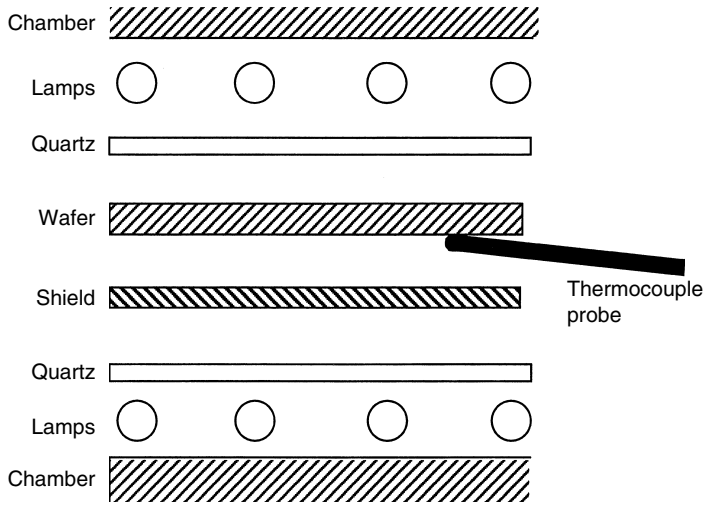


FIGURE 11.45 The combination of a radiation shield approach with a TC probe yields excellent emissivity-independence. In the steady-state, the probe is in a near-isothermal environment.

from both above and below by lamps, as illustrated in Figure 11.45, the probe sits in a near-isothermal environment and achieves excellent emissivity-independence. Figure 11.46a shows the steady-state temperatures expected when coated wafers are processed under conditions where the probe reaches a temperature of 1000°C. The wafer's backsides are coated with various thicknesses of polysilicon on top of

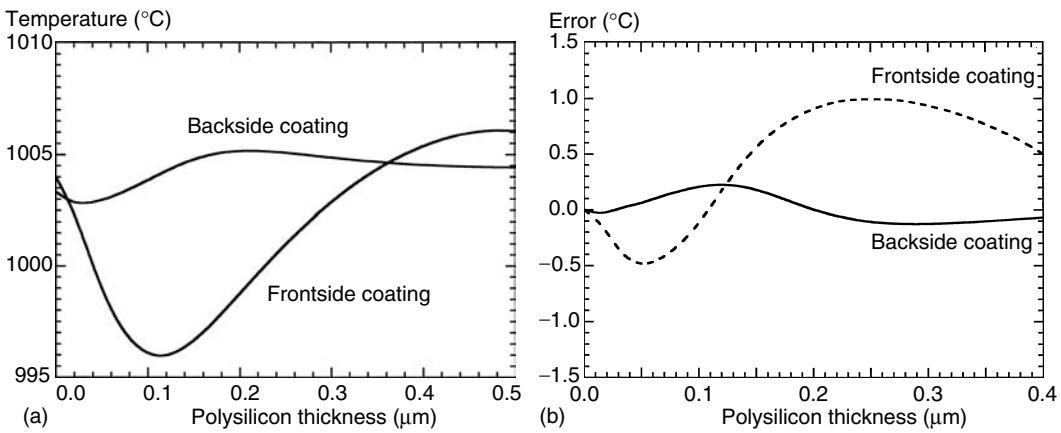


FIGURE 11.46 (a) The effect of coating film thickness on the wafer temperature reached during heating under the control of a TC probe in the gap between the wafer and a radiation shield. The calculations show the effects of polysilicon film thickness for a wafer that has a coating with a polysilicon film over 0.2 μm of silicon dioxide. The behavior is shown for the cases where the coating is on either the front or the back of the wafer. The probe is assumed to be at 1000°C. The probe and the shield are both assumed to be opaque and spectrally grey, with emissivities of 0.8. The calculations show that the approach virtually eliminates the effect of the backside coatings, and the effect of variations in the frontside coating is reduced relative to that shown in Figure 11.42a (b) the effect of film thickness changes on the wafer temperature. The calculations demonstrate the impact of a 10% increase in the film thickness of a polysilicon film for a wafer coated with polysilicon on top of 0.2 μm of oxide. The results show that this approach is very robust with respect to variations in wafer coatings.

a 0.2-μm thick oxide layer. As before, the calculations did not include the dynamics and the quartz tube temperature was fixed at 400°C. The results show a 2°C temperature range when the coating is on the wafer backside, and a 10°C range when the coating is on the wafer frontside. The method reduces the impact of frontside coating changes compared to the case where the shield temperature is controlled. Figure 11.46b shows that the +10% changes in frontside polysilicon film thickness introduce temperature changes $< \pm 1$°C, providing excellent immunity to film thickness fluctuations. In practice, the TC probe is placed in contact with the wafer backside, so that the link between the temperature of the probe and the wafer is even closer than that between the temperature of the shield and the wafer. This should reduce the effects of frontside coatings, fluctuations in the lamp zone settings, and the condition of the quartz tube.

11.2.7.10 Summary of RTP Temperature Measurement Capabilities

Table 11.16 summarizes the main strengths and weaknesses of current RTP temperature control methods and lists some of the practical considerations involved in their deployment in RTP systems. The table demonstrates that several approaches can give very repeatable results, and the practical choice may be dictated by other factors, including system cost and reliability. It is also important to realize that as RTP temperature measurement and control have improved, the influence of non-thermal factors on perceived process repeatability is becoming more significant. Table 11.17 lists some of these factors [6,142–144]. Furthermore, the complex interplay between various physical processes often makes process results sensitive to the whole temperature–time cycle, rather than just a peak or a steady-state temperature [145]. In order to make the best use of any given process monitor, a good understanding of its kinetics is necessary. Because of these factors, it can be a significant challenge to establish the ultimate repeatability of the RTP temperature measurement system. It has even been proposed that operating an RTP system under stabilized open-loop lamp power control conditions can provide an extremely repeatable temperature standard, which might even be better than many familiar process monitors [132,146]. The next stage in sophistication is to move from very repeatable temperature performance to high absolute accuracy, where the tool's temperature readings can be linked back to absolute standards [104]. In the past, various systematic temperature calibration errors have resulted in different process results from different RTP tools running nominally the same process [139,147]. Although, this is not necessarily very important in day-to-day manufacturing, improvements in absolute accuracy could provide advantages when comparing process results from different tools, especially if those tools are at different sites or are made by different manufacturers.

11.2.8 Process Uniformity Control in RTP

The RTP temperature control problem arises from three elements, the uniformity, repeatability, and independence from the type of wafer. The uniformity part of the problem is probably the one which has the most impact on the basic layout of an RTP system design. In conventional hot-wall furnaces, the entire reaction chamber, which contains a large batch of wafers, is kept at one temperature. This gives good temperature uniformity in the steady-state because all the elements inside the furnace reach the same temperature eventually. However, the heating-rates are inherently very low, and if one tries to heat the wafers faster by ramping the furnace-wall temperature faster, or by pushing the wafers into the furnace faster, then the temperature uniformity becomes worse. As a result, conventional furnace systems cannot simultaneously provide uniform processing and meet tight restrictions on thermal budget. The move to larger wafer sizes also fundamentally clashes with this technology, because conventional hot-wall systems heat wafers at their edges much more than at their centers. The heating rate can be increased and made more uniform by increasing the spacing between the wafers in the batch, but this decreases the number of wafers which can be loaded in the furnace. In contrast, for RTP systems, the scale-up of equipment to handle larger wafer sizes is relatively straightforward, because the heaters face the surfaces of the wafer. The design problem may even become simpler, because some of the more difficult aspects

TABLE 11.16 Summary of Relative Merits of Current Temperature Control Methods

Method	Principle	Advantages	Disadvantages	Comparative Estimate of Temperature Range as Films Vary
Pyrometry	Measurement of intensity of thermal radiation over a narrow range of wavelengths	Closed-loop control of wafer temperature Simple Fast response Multi-channel systems allow dynamic uniformity control	Requires sophisticated emissivity correction to achieve good results Requires excellent rejection of interfering radiation sources Signal level limits minimum temperature Wafer transparency can make low-temperature measurements difficult	Estimates for 2.7 μm pyrometry No ε-correct, $R_c = 0$: ± 180°C No ε-correct, $R_c = 0.95$: ± 18°C 10% effect, $R_c = 0$: ± 83°C 10% effect, $R_c = 0.95$: ± 7°C In situ correction: ± 1°C
Lamp intensity control	Repeatable cycle of lamp intensity	Simple Insensitive to minor fluctuations in wafer structure	No feedback Wafer effects can be large—requires calibration Requires excellent stability in lamp power supply and system facilities Aging of system components and deterioration of system optics affect results	Coating effect: ± 25°C 10% effect: ± 4°C
Susceptors and hot-plates	Repeatable susceptor/hot-plate temperature	Simple Insensitive to minor fluctuations in wafer structure	No feedback Wafer effects can be large—requires calibration Limited possibilities for shape of thermal cycle and response time Sensitive to gap between wafer and hot surface, gas type and pressure	Coating effect: ± 25°C 10% effect: ± 4°C
Contact thermocouple	Repeatable thermocouple temperature	Simple Insensitive to minor fluctuations in wafer structure Feedback from thermocouple indicates process conditions	Feedback is not directly from the wafer itself Wafer effects can be large—requires calibration Differences in transient thermal response of thermocouple and wafer have to be taken into account in recipe set-up	Backside coating effect: ± 27°C 10% backside effect: ± 3°C Frontside coating effect: ± 6°C 10% frontside effect: ± 1°C

(*continued*)

TABLE 11.16 ***(Continued)***

Method	Principle	Advantages	Disadvantages	Comparative Estimate of Temperature Range as Films Vary
Radiation shield	Repeatable shield temperature	Simple Very insensitive to changes in wafer backside coatings Shield can improve process uniformity	Feedback is not from wafer itself, changes in thermal conditions, e.g., due to component aging, can affect results Can be affected by changes in wafer topside films Differences in transient thermal response of shield and wafer have to be taken into account in recipe set-up Extra thermal mass of shield reduces maximum ramp rates and system throughput	Backside coating effect: ±1.4°C 10% backside effect: ±0.2°C Frontside coating effect: ±10°C 10% frontside effect: ±1.5°C
Contact thermocouple between radiation shield and wafer	Repeatable thermocouple temperature, thermocouple in cavity between wafer and shield	Simple Very insensitive to changes in wafer backside coatings Feedback from thermocouple reduces impact of wafer frontside conditions Shield can improve process uniformity	Feedback is not directly from wafer itself, radical changes in wafer topside films can still have some effect Differences in transient thermal response of thermocouple and wafer have to be taken into account in recipe set-up Extra thermal mass of shield reduces maximum ramp rates and system throughput	Backside coating effect: ±1.2°C 10% backside effect: ±0.2°C Frontside coating effect: ±5°C 10% frontside effect: ±1°C

The right-hand column aims to quantify the capabilities of the techniques by comparing theoretical predictions of the performance obtained by methods described in the text. The temperature ranges refer to the range of wafer temperatures obtained for wafers that have a coating with a polysilicon film on top of a 0.2-μm thick oxide layer. The range of temperatures describes the variation expected as the polysilicon film thickness varies from 0 to 0.5 μm. The "10% effect" refers to the error expected when the temperature control technique has been calibrated for a given polysilicon thickness, but some disturbance causes the actual film to be 10% thicker. Some techniques are sensitive to variations in coatings on the front of the wafer as well as those on the back. In these cases separate values have been given for the effects of coatings on the two surfaces. (R_c, Chamber reflectivity.)

TABLE 11.17 Process Control Problems Not Linked to Temperature Error

Type of Process	Problem
Silicide formation	Process ambient contamination, especially O_2 and H_2O
	Film thickness variation
	Contamination between metal film and silicon, especially oxide
	Film impurities
	Film stress
	Metrology problems
Rapid thermal anneal	Ion implantation dose variations
	Screen oxide thickness variations
	Variations in ambient gas composition
	Metrology problems
Rapid thermal oxidation	Surface preparation problems (microroughness, chemical contamination, wafer "quality" and particles)
	Native oxide variations
	Process ambient contamination (especially H_2O)
	Atmospheric pressure fluctuations
	Post-processing surface contamination
	Metrology problems

Many of these effects can be confused with temperature calibration or temperature uniformity problems.

arise from heat loss from the edges, which occupy a smaller fraction of the wafer area as the wafer size rises [13].

The RTP systems shown in Figure 11.1 all handle uniformity control in different ways. The following sections review the physics underlying the uniformity optimization problem and the hardware implications, and consider some practical methods for process optimization. Although RTP process uniformity is often regarded as being limited by the wafer temperature uniformity, other factors can become more significant. Table 11.17 shows some non-thermal process factors that can affect the uniformity in several important RTP applications. In some applications, gas flow patterns can directly affect process uniformity [148]. This is especially relevant when a reaction tends to deplete a component from the gas and produce a concentration gradient over the surface of the wafer. This factor is often important in RTCVD systems.

11.2.8.1 Physics of the Temperature Uniformity Problem

11.2.8.1.1 Steady-State Non-Uniformity

In the steady-state part of a heating cycle, temperature uniformity depends on how the balance between radiant power delivered and thermal loss varies with position on the wafer. A simple model can be used to illustrate the physical phenomena involved. For an infinite sheet of material with a thickness D and thermal conductivity K, the $1D$ steady-state problem for heat flow in the x-direction can be summarized by the equation,

$$KD\frac{\mathrm{d}^2T(x)}{\mathrm{d}x^2} + \eta(x)P(x) - H_{\mathrm{eff}}(x)\sigma T(x)^4 = 0, \tag{11.37}$$

where $\eta(x)$ is the power coupling efficiency, $P(x)$ is the power density distribution, σ is the Stefan–Boltzmann constant, $T(x)$ is the wafer temperature, and $H_{\mathrm{eff}}(x)$ describes the efficiency with which power is lost by thermal radiation [149]. If D or K is very small, then the temperature is defined by the power balance in Equation 11.25, where η, P, H_{eff}, and T take values appropriate for each position on the wafer. In reality, silicon is a good conductor of heat and lateral thermal conduction within the wafer smoothes out the temperature profile, reducing the non-uniformity. The smoothing effect becomes stronger as the length scale considered becomes smaller. For a $1D$ sinusoidal modulation of the lamp radiation flux incident on the wafer, where the modulation is small compared to the average incident power, we can

analyze the problem with a linearized version of Equation 11.37,

$$KD\frac{d^2\Delta T(x)}{dx^2} + \eta\Delta P \sin\left(\frac{2\pi x}{L}\right) - 4T_0^3 H_{\text{eff}}\sigma\Delta T(x) = 0, \tag{11.38}$$

where $\Delta P \sin(2\pi x/L)$ is the sinusoidal power density modulation, with period L, $\Delta T(x)$ is the resulting modulation of the temperature and T_0 is the mean temperature, given by Equation 11.25. Solution of Equation 11.38 gives

$$\Delta T(x) = \frac{\eta\Delta P}{4T_0^3\sigma H_{\text{eff}} + KD(2\pi/L)^2}\sin\left(\frac{2\pi x}{L}\right). \tag{11.39}$$

This expression shows that as the length scale, L, decreases, the non-uniformity reduces [88,149]. Also, as the wafer thickness or the thermal conductivity increases, the non-uniformity decreases. Figure 11.47 shows the predicted magnitude of the temperature modulation resulting from a 1% sinusoidal modulation of the incident power required to keep the wafer various temperatures, for a range of length scales. The wafer is 725 μm thick, η and H_{eff} were taken as unity and the thermal conductivity of silicon was taken from Ref. [150]. In the limit of very large length scales, the modulation approaches the value for no conduction within the wafer. The non-uniformity increases rapidly with temperature, for all length scales, because ΔP rises with T_0^4. The thermal conductivity of silicon also decreases rapidly as the temperature rises, which makes non-uniformity problems worse [150]. For length scales below ~3 mm, non-uniformity is negligible for power changes <10%.

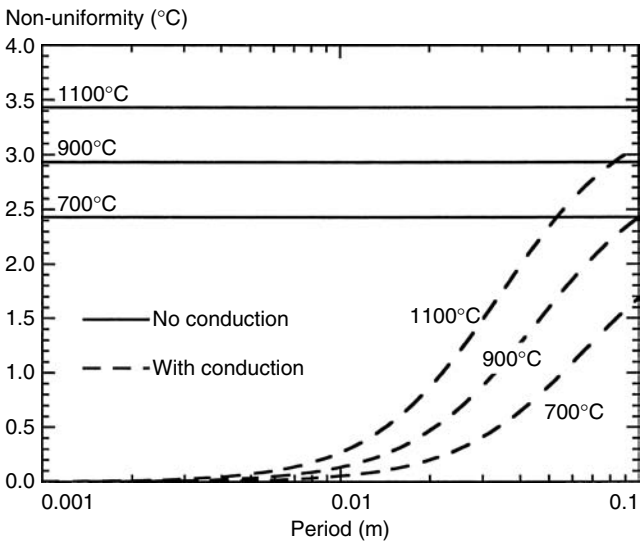


FIGURE 11.47 The effect of a spatial modulation of incident power density on wafer temperature uniformity. The curves show predictions of the magnitude of the temperature non-uniformity caused by a 1% sinusoidal modulation in the input power density for a 725-μm thick silicon wafer at 700, 900, and 1100°C. The graph shows the non-uniformity as a function of the period of the modulation. The results expected for a case where there is no thermal conduction within the substrate are included for comparison. Thermal conduction within the silicon is very effective at suppressing temperature non-uniformity at length scales <~3 mm. As the length scale increases, the non-uniformity approaches the limit of the case without conduction.

11.2.8.1.2 Transient Non-Uniformity

During the ramp-up and down in temperature, the non-uniformity problem is rather different from that in the steady-state, because some of the energy delivered to the wafer is used to increase the wafer temperature, rather than to just compensate for heat loss. This changes the spatial power distribution required for temperature uniformity. For example, at the very first moment when power is applied, there is very little heat loss from the wafer, because it is very close to the chamber temperature. In this condition, temperature uniformity only requires uniformity in the lamp power absorption across the wafer. As the temperature rises, and spatially inhomogeneous heat loss appears, for example, because the wafer edge loses more heat than the center, the flux distribution required for uniformity changes. As a result, the illumination distribution required for a uniform temperature evolves as the wafer heats up. The transition from the transient condition to the steady-state can take quite some time because of the time constant of the wafer shown in Figure 11.23b. Transient temperature non-uniformity can have a large impact on short processes, such as spike anneals of ion implantation damage.

11.2.8.1.3 Temperature Difference across the Wafer Thickness

The previous discussion has implicitly assumed that the temperature of the front and back surfaces of the wafer is equal. This is close to reality because the wafer is thin and silicon has a high thermal conductivity. This ensures that the temperature difference is small, and also that the time lag between the temperature response of the two surfaces is <0.1 s. In RTP systems, which heat the wafer from both sides, the temperature differences are automatically very small. In systems, which heat the wafer from one side, the temperature difference between the heated and unheated sides, ΔT_d, during ramp-up can be estimated from the simple approximation provided by Vandenabeele,

$$\Delta T_d \cong \frac{c\rho D^2}{2K}\frac{dT}{dt} + \frac{H_{eff,u}\sigma D T^4}{K}, \tag{11.40}$$

where dT/dt is the ramp rate, $H_{eff,u}$ is the effective total emissivity of the unheated side of the wafer and the other symbols have the meanings given before [149]. The temperature difference is greatest at the moment just at the end of the ramp-up. In the steady-state, $dT/dt=0$, and the temperature difference reduces significantly. If $H_{eff,u}$ is made small, for example, by placing a highly reflecting plate next to the unilluminated side of the wafer, then the temperature drop becomes very small. For a ramp to 1000°C at 100°C/s, Equation 11.40 predicts that the temperature difference across a 725-μm thick plain silicon wafer in a "black" environment, where $H_{eff,u}$ is 0.7, is $\sim$5.3°C at the end of the ramp, and it drops to 3°C in the steady-state. If a reflective plate is used to reduce $H_{eff,u}$ to 0.2, this temperature decreases to <1°C in the steady-state.

11.2.8.2 Configurations for Uniformity

Traditional RTP systems, illustrated in Figure 11.1a, used rectangular reaction chambers combined with linear lamp arrays, with different groups of lamps gathered together as zones. Further improvements in temperature uniformity are possible by introducing rotation [5]. Wafer rotation is often combined with axisymmetric lamp arrays as shown in Figure 11.1b. Systems with only one lamp, such as the arc lamp approach illustrated in Figure 11.1c, have poor ability for uniformity tuning, and have been largely replaced by multi-zone systems. Rapid thermal processing systems that use a hot-plate approach, like that illustrated in Figure 11.1d, can incorporate a multi-zone heater to tune uniformity of the hot-plate, but the basic assumption is that a uniform hot-plate will produce a uniform wafer temperature distribution. The type of system shown in Figure 11.1e is essentially a hot-wall system, and the heat transfer is similar to that in a furnace.

Table 11.18 summarizes some generic causes of non-uniformity for typical, lamp-based RTP systems and some of the system features and operating procedures which have been used to address these issues. Some effects arise from the system design limitations, and others derive from the wafers themselves.

TABLE 11.18 Origins of Temperature Non-Uniformity Problems in RTP and Possible Solutions

Type of Problem	Origin	Solution
Lamp irradiation pattern	Inhomogeneous incident flux from lamps	Multi-zone illuminator Optimized lamp illumination profiles Wafer rotation Radiation shield/susceptor Reflective chamber design MIMO control/model-based control *Zone tuning*
Dynamic change of power distribution needed for uniformity	Progressive evolution of requirements for illumination distribution as wafer heats up	MIMO control/model-based control *Zones tuned differently in several recipe blocks e.g., ramp stage and steady-state stage*
Edge overheating in ramp stage	"Photon-box" effect in a reflective cavity results in increase in incident power at wafer edge Increased surface to volume ratio at wafer edge Increased surface to volume ratio at wafer edge	Reflective chamber design Slip-free ring MIMO control/model-based control *Reduced ramp rate* *Separate optimization of ramp stage and steady-state zone settings*
Edge losses	Increased surface to volume ratio at wafer edge Difference in chamber reflectivity at wafer edge Gas flow impinging on wafer edge	Slip-free ring MIMO control/model-based control *Zone tuning* *Reduce gas flow rate*
Gas flow effects	Convective cooling from gas impinging on wafer	Gas inlet design (including showerhead concepts) Wafer rotation Slip-free ring *Reduce gas flow rate*
Quartz heating	Progressive changes in heat input to wafer from quartz windows and tubes which heat up during processing, changing wafer temperature distribution	Effective quartz window/tube cooling methods MIMO control/model-based control *Preheat quartz with lamps and dummy wafers* *Maintain correct facilities, especially compressed air and water cooling*

Effect of chamber component changes on illumination pattern	Chamber contamination from outgassing wafers, especially on quartz isolation tubes and windows, changes heat transfer conditions	MIMO control Removable quartz chamber liner which can be cleaned *Clean chamber, tube or window*
	Lamp aging affects illumination profile	*Replace lamps*
	Changes in slip-free rings (e.g., progressive oxidation, progressive mechanical deformation) affect heat transfer, especially at wafer edge	*Stabilize slip-free ring properties by pre-processing* *Clean or replace slip-free ring*
	Lamp reflector changes	*Clean or replace lamp reflectors*
Unexpected changes in illumination pattern/heat losses	Line voltage fluctuations affecting different zones in different ways	MIMO control
	Facilities fluctuations affecting heat loss mechanisms	Closed-loop power control to lamp zones *Maintain correct voltage supply, via line-voltage regulator if necessary* *Maintain correct facilities, especially compressed air cooling, water cooling and exhaust*
Mechanical misalignment	Wafer position/tilt error affects illumination profile and heat losses	*Ensure wafer and wafer support loaded correctly in RTP chamber*
Pin or ring support effects	Local heat loss and thermal mass effects from support pins or support ring	Double-sided heating Optimized pin shape
Pattern effects	Impact of patterned coatings on power coupling and heat loss (possible on both front and back of wafer)	See separate table

The items in italics refer to solutions that can be implemented by the user, the other solutions must be addressed in the system design.

For example, the "pattern effect" is caused by variations of thermal radiative properties across the wafer surface, and will be discussed in more detail below.

The stringent demands on the uniformity of the power balance discussed above mean that the RTP heating system has to be designed with great care. The problem of designing a suitable heater is not necessarily the same as that of designing a very uniform illuminator. Uniformity optimization requires controllability as well as uniformity, where controllability refers to the ability to locally affect the temperature in selected regions of a wafer relative to another region, for example, to counteract non-uniformity in the power loss distribution [13]. A heater with several zones of lamps, which can be adjusted independently, can provide uniformity despite the differing power delivery requirements for various process recipes, wafer types, and gas flow conditions. The multi-zone structure gives the process engineer a way to optimize process uniformity, and it can also be used to provide automatic uniformity optimization, in real time, via automatic control. Uniformity optimization techniques will be discussed in more detail below.

A large improvement in wafer uniformity can be obtained through wafer rotation, which smoothes out non-uniformities in power delivery and heat loss [13]. Figure 11.48 shows how wafer rotation reduces the impact of azimuthal non-uniformities in the heat transfer conditions. In this example, it is assumed that the lamp power distribution is azimuthally non-uniform and that for a quarter of the wafer, the lamp flux is 10% lower than for the rest. For a wafer at 1100°C, this segment would be 36°C cooler than the rest of the wafer. A simple zero-dimensional model, developed from Equation 11.24, was used to evaluate the effect of rotation at various rates on this non-uniformity, by making power input a function of time during the rotation cycle. The wafer was 725 μm thick, and η and H_{eff} were taken to be unity. Figure 11.48 shows that the temperature at a point on the wafer oscillates, and that as the rotation rate rises, the amplitude of this oscillation decreases. Figure 11.49 shows that most of the benefit of rotation is achieved at rotation rates below $\sim$30 rpm, above which there are diminishing returns in increasing the rate. The rotation reduces a 36°C temperature non-uniformity in a particular region of the wafer to an

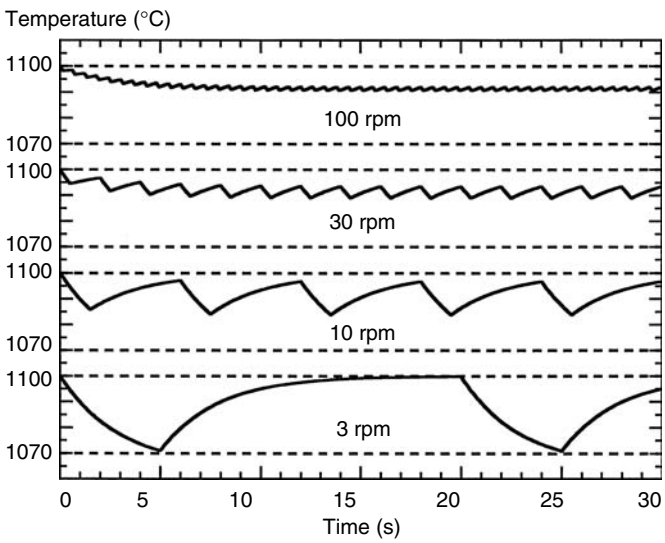


FIGURE 11.48 The effect of wafer rotation on the thermal response of a wafer. In the example shown it is assumed that the wafer sits in a chamber where one quarter of the wafer is illuminated with 10% less lamp power than the rest of the wafer. The curves show predictions of the transient temperature response of a point on the wafer that starts off in the brighter part of the chamber at 1100°C and then rotates through the darker segment. As the rotation rate increases, the thermal mass of the wafer damps out the temperature oscillation.

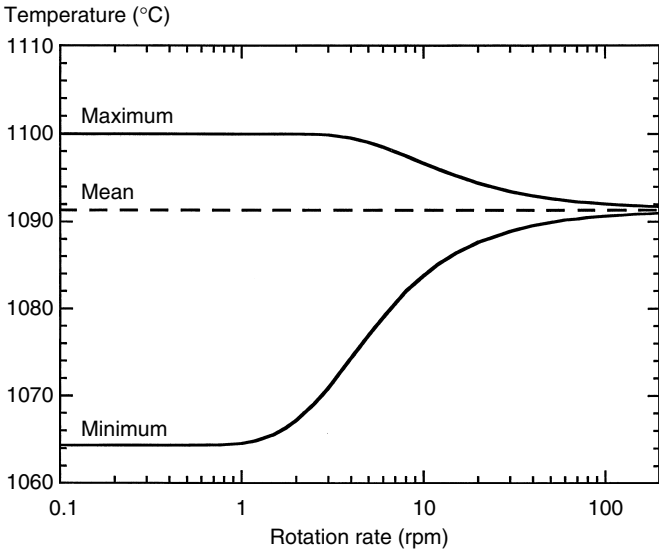


FIGURE 11.49 The effect of rotation rate on temperature non-uniformity. The results show the maximum and minimum temperatures experienced by a point on the rotating wafer described in Figure 11.48. As the rotation rate rises, the temperature oscillation decreases and approaches the average value. The figure demonstrates that most of the benefit of rotation is achieved at rates $< \sim 30$ rpm.

axisymmetric 9°C drop, which could be tuned out by adjusting lamp zone settings. Effects of azimuthal non-uniformities in heat loss can also be attenuated and made axisymmetric. Rotation has been incorporated into most advanced RTP tools [5–8].

One important aspect of the uniformity problem in RTP arises from the geometric differences between the central region of the wafer and the edge [58,79,80,82]. Since the early days of RTP, it has been known that the lamp power coupling and the heat losses at the edge tend to be quite different from those nearer to the wafer center, and that this could produce large transient and steady-state temperature non-uniformities [58,80,151,152]. The temperature non-uniformity could also result in stresses that produced slip lines and wafer distortion [153–155]. Various approaches have been developed to minimize the effect of the edge, including the use of slip-free rings and special reflectors [79,80,153]. The ability to minimize transient temperature non-uniformity by separate adjustment of zone ratios for ramp-up and for steady-state stages of heating cycles has also greatly reduced the temperature non-uniformities. This approach has now evolved to permit dynamic optimization of temperature uniformity throughout the heating cycle.

11.2.8.3 Uniformity Optimization Procedures

Reaching the optimum process uniformity is an important aspect of the process engineer's work, since the device yield can be degraded by poor uniformity. Typically, adjustment of zone settings is performed by processing monitor wafers and measuring their uniformity. Monitor processes are often based on thermal oxidation, annealing, and electrical activation of ion-implanted layers or on metal silicide formation. Some non-thermal factors, which can impact these processes, are discussed in Table 11.17.

Although it is possible to optimize process uniformity by trial-and-error adjustments, systematic approaches are often useful. One method involves the generation of a "response surface" or "gain matrix" which links the process uniformity and the zone settings [156,157]. This involves taking a baseline set of zone settings that are known to give reasonable uniformity and then processing several monitor wafers in a series of experiments where each lamp zone, one at a time, is increased by a known fraction relative

to the other settings. For example, in a six-zone RTP system, seven experiments could be conducted, a baseline case and six other cases in which each zone setting is increased by 10% relative to its baseline value. The point-by-point differences between the wafer metrology data for each of these "+10%" runs and the baseline can then be calculated and scaled with respect to the zone variation used. The resulting data can be organized as a gain matrix, **G**, defined by the equation

$$\mathbf{M} = \mathbf{GZ}, \tag{11.41}$$

where **M** is a vector containing the metrology results at each point on a wafer map and **Z** is the vector of lamp zone powers. This equation can be used to predict the metrology data that will result from any given change in the zone settings. An automatic minimization procedure can then be used to predict the optimal zone settings. The approach has been tested and found to provide a robust method of process optimization. Although several monitor wafers are consumed in generating the gain matrix, the gain matrix generated on one system can often be used on other systems. Since RTP processes are usually short in duration, it is often best to optimize the temperature uniformity in the ramp-up stage as well as in the steady-state part of the recipe. This can be done by creating a recipe which only has the ramp to the process temperature, and using a gain-matrix method to optimize the zone settings for this part of the process. It is also possible to derive gain matrix information through theoretical simulations of the effects of changes in lamp zone powers. This approach can be calibrated with empirical data to provide a model that can be used to predict optimized lamp zone power distributions [93,158].

11.2.8.4 Dynamic Temperature Uniformity Control

Rapid thermal processing technology has taken a considerable step forward with the introduction of systems that can provide uniform heating of the wafer throughout the heating cycle. In early RTP systems, a pyrometer would monitor the temperature at one location on the wafer and provide feedback to the power control algorithm. The ratio of the power delivered to any lamp zone was kept fixed in any block of a processing recipe, so all the zones were "slaved" together to the one feedback signal. However, it is well-known that the spatial distribution of heat loss from the wafer varies dynamically as the wafer temperature changes, and hence a single set of fixed zone ratios cannot provide a uniform heating condition throughout the heating cycle [159]. Significant improvements in the dynamic uniformity could be made by dividing the recipe into various "blocks" of time, and optimizing the zone ratios for uniformity in each recipe block [157]. For example, one set of zone ratios could be used in a ramp-up stage, and a second set used in the steady-state stage, as mentioned above. Although this method has met with considerable success, it is quite time consuming to optimize recipes and typically several monitor wafers have to be consumed in order to obtain optimal uniformity. Furthermore, it may still be necessary to warm up the process chamber by running several dummy wafers at the start of a lot, in order to reach stable heat transfer conditions. As a result, there has been great interest in development of approaches that automatically provide dynamic temperature uniformity throughout the heating cycle and that simplify recipe set-up. Two approaches have evolved to meet this need. One approach is the use of model-based control, where the system controller dynamically adjusts the lamp power settings in order to take into account the variation in heat transfer conditions throughout the recipe. Such approaches can be implemented through the use of a sophisticated physical model of the system that can predict how energy is transferred to the wafer from the lamps and the heat transfer between the wafer and the chamber components, including aspects such as the warm-up of quartz windows [160]. When combined with high-stability lamp power supplies this approach has the merit of providing a dynamic temperature control, a simple uniformity tuning procedure, and a reliable system. The challenge for a purely model-based approach arises if the wafer properties vary greatly, since the model must then adapt to match any corresponding variation in heat transfer conditions. As a result, model-based control schemes work best in chamber configurations with a high degree of inherent uniformity, and the RTP system design must be optimized to make non-uniformity as insensitive to wafer properties as possible.

Another approach for dynamic uniformity control involves adding extra temperature sensors that can monitor the wafer temperature at multiple positions on the wafer. These inputs can be provided to an MIMO control system that uses the sensor information to dynamically determine lamp power settings [13]. There have been many theoretical and experimental studies of how to design an MIMO controller, and various control algorithms have been proposed [6–8,87,161,162]. In principle, MIMO control schemes provide closed-loop control of the temperature distribution, which should allow them to adapt the heating conditions to take account of variations in wafer properties or heat transfer conditions. The challenge for MIMO control systems falls in the need to provide extremely close matching of the readings of several temperature sensors. Since the total range of temperatures across the wafer needs to be below $\sim \pm 1°C$, it is essential for the sensors to be matched to well below this limit. It is quite difficult to meet this objective dynamically throughout the entire temperature range of a recipe. Furthermore, all the temperature readings must be correct regardless of the spectral emissivity of the wafer at the location being monitored by the sensor. Such challenges must be met, otherwise the MIMO control scheme can actually degrade the wafer uniformity.

11.2.8.5 The Effects of Coatings and Patterns on RTP Uniformity

11.2.8.5.1 The Effect of Uniform Coatings on Temperature Uniformity

Process recipes are usually established using plain monitor wafers, but product wafers are often coated with films that change the thermal radiative properties and affect heat transfer [14,64,72–75,163,164]. Process uniformity can be affected by this kind of change. The effect can be evaluated by considering uniformity optimization problem depicted in Figure 11.50. The absolute temperatures at two points on the wafer, A and B, are given by T_A and T_B, respectively. The RTP system has two zones, and when zone 1 delivers a power P, a fraction η_{A1} is absorbed at point A and η_{B1} is absorbed at B. Zone 2 is set to emit a fraction Z of the power from zone 1, and the fractions absorbed at points A and B are η_{A2} and η_{B2},

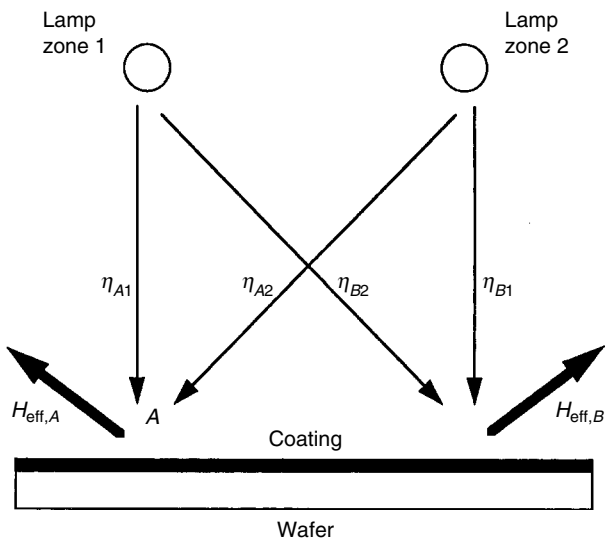


FIGURE 11.50 Illustration of the effects of a coating that changes the lamp power coupling and heat loss properties differently at two places, A and B on a wafer. If the coating is different at points A and B, a "pattern effect" can be observed. Even if the coating is the same, chamber and lamp zone configuration can result in the coating having a different effect at A and B. Either way, the temperature uniformity can be altered.

respectively. From Equation 11.25,

$$\frac{T_A}{T_B} = \left(\frac{\eta_{A1} + Z\eta_{A2}}{\eta_{B1} + Z\eta_{B2}} \frac{H_{\text{eff},B}}{H_{\text{eff},A}}\right)^{1/4}, \tag{11.42}$$

where $H_{\text{eff},A}$ and $H_{\text{eff},B}$ are the local values for the heat loss efficiency at points A and B, respectively. This equation can be used to deduce the lamp zone ratio, Z_{u}, required to make $T_A = T_B$,

$$Z_{\text{u}} = \frac{\eta_{B1} H_{\text{eff},A} - \eta_{A1} H_{\text{eff},B}}{\eta_{A2} H_{\text{eff},B} - \eta_{B2} H_{\text{eff},A}}. \tag{11.43}$$

If the thermal radiative properties of the wafer change in such a way that the change in H_{eff} or η at points A and B is different, then the zone ratio Z_{u} will no longer give temperature uniformity. This is clearly the case if the wafer is patterned in a manner that makes H_{eff} or η different at A and B. However, it is also possible even with a uniform coating, because the chamber and lamp configuration can influence the values of H_{eff} or η in a different way at different positions on the wafer. One simple example is in a double-side-heated RTP system, where zone 1 corresponds to lamps above the wafer, and zone 2 corresponds to the lamps below the wafer. The system might be optimized using a plain silicon wafer, but when a wafer with a backside coating is processed there is a change in the power coupling, η_{A2} and η_{B2}, for the lamps below the wafer, but η_{A1} and η_{B1} remain the same. As a result, the zone ratio required for uniformity would no longer equal Z_{u}.

Other complications can arise from the use of heavily doped substrates on product wafers when the monitor wafers are lightly doped. The dynamics of the heating cycles change significantly when the substrate doping changes, and there could be an impact on process results. In principle, such problems can be alleviated by including the wafer properties in a model-based control scheme or by closed-loop control of the wafer temperature distribution rather than just one point on the wafer.

11.2.8.5.2 The Effect of Non-Uniform Coatings: The Pattern Effect

When the coating on the wafer surface is non-uniform, for example, because of a device pattern on the wafer front, or a non-uniform film deposition on the back of the wafer, there will be some non-uniformity in the temperature distribution [58,72,80,87,94,136,165–178]. This pattern effect will usually be small for features less than ~3 mm in size because of the effect of thermal conduction. Individual device structures are typically much too small to have any thermal impact, but regions such as the unpatterned zone at the edge of the wafer, where the die pattern has not been imprinted, can cause uniformity problems. Other large regions such as global alignment marks are also known to cause pattern effect problems [175]. Furthermore, as die sizes have increased and system-on-chip technologies have become important, there are many die that may have rather large variations in optical properties even for different circuit blocks within the die. Usually, such non-uniformities cannot be removed by lamp zone adjustments, regardless of the control scheme, because the irradiation pattern from any given lamp is usually too broad to compensate for such local non-uniformities. Furthermore, the non-uniformity is not relieved by wafer rotation, because it originates from features on the wafer surface. Indeed, problems can be compounded if the wafer is rotating and a MIMO control scheme is used, because the non-uniform heating disturbs the temperature sensor readings but the lamps are unable to compensate for it [175]. Wafer backsides are also sometimes coated with non-uniform films that can form a pattern that affects the temperature uniformity.

Various schemes have been presented for dealing with pattern effect issues in RTP. As uniformity of RTP processes has improved to the point where 1-sigma temperature variation is only ~0.5°C, the non-uniformity from pattern effects has grown in significance, and indeed it may be the dominant non-uniformity in production [176,177,179]. Pattern effects have been associated with both process non-uniformity and with introducing stresses within wafers that can introduce defects and even lead to problems in alignment of patterns [177]. Table 11.19 discusses the relative merits of various schemes

TABLE 11.19 Methods for Controlling the "Pattern Effect" in RTP Systems

Method	Principle	Comment
Minimize patterns with length scale > ~5 mm on wafers	Side-steps the problem	One example is to print the die pattern right out to the edge of the wafer, eliminating a large unpatterned region around the wafer edge, which can have different optical properties to those of the die itself
MIMO control	Uses temperature feedback to correct for pattern-induced non-uniformity	This is only likely to help with problem of an unpatterned edge region in systems where the wafer is rotating, so that a sensor near the wafer edge can give appropriate feedback
Illuminate only the unpatterned side of the wafer	Reduces the effects of variations in lamp power coupling	Variations in integrated emissivity (causing non-uniform heat loss) can still have an impact There may be non-uniform deposition patterns with large length scales on the back of the wafer, which means that it is not always clear which wafer surface has the more significant "pattern"
Arrange for the patterned surface to face a broad-band reflector	Reducing the heat loss from the patterned surface minimizes the impact of variations in integrated emissivity	There may be non-uniform deposition patterns with large length scales on the back of the wafer, which means that it is not always clear which wafer surface has the more significant "pattern"
Double-side heating	Splits the problem of lamp power coupling variations between the two sides of the wafer	
Thermal equilibrium	Thermal radiative properties do not matter for a system in thermal equilibrium	Only absolutely true for wafer in a hot-wall environment at steady-state. However, the nearer the heat source temperature is to the wafer temperature, the less important pattern effects are
Hot-plates, susceptors and radiation shields	Pseudo-equilibrium conditions for patterned wafer surfaces facing the local "hot-wall"	Works on the same principle as the hot-wall case. The wafer could be placed between two "hot-plates" to maximize the effectiveness
Reduced ramp rate	Reduces the impact of lamp power coupling variations during the ramp stage	This approach only helps reduce transient non-uniformity, but this can be important

for reducing the pattern effect in RTP. A dominant factor is the choice of heating configuration, and Figure 11.51 illustrates the impact of this choice on the pattern effect during an ion implantation annealing process [179]. In this experiment, the wafer was patterned with a checkerboard pattern on its frontside and the metrology was performed on the backside, which had received the implant. The RTP process was an 1100°C spike anneal of wafers implanted with 10^{15} As/cm^2 at 1 keV. The results show the sheet resistance variation observed on a linear scan across part of the wafer. The temperature scale is derived from the 1 Ω/(sq. °C) sensitivity of the process. The simplest approach for reducing the pattern effect, while maintaining full flexibility in temperature cycles, is through the use of dual-sided irradiation, which can reduce the pattern effect by 50% relative to a frontside heating approach. Complete elimination of the pattern effect is possible through the use of the "Hot Shielding" approach, which retains most of the flexibility in thermal profiles [174,178]. In this method, a silicon plate is interposed between the lamps and the patterned surface of the device wafer. In a dual-sided heating system,

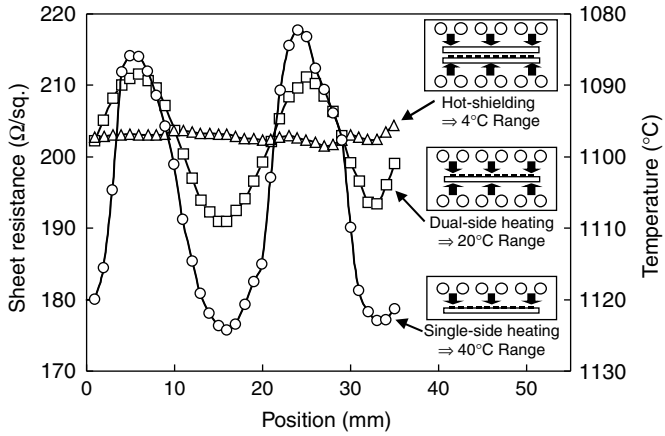


FIGURE 11.51 The effects of RTP heating configuration on pattern effects during an 1100°C spike anneal of wafers implanted with 10^{15} As/cm^2 at 1 keV. (Reprinted From MacKnight, R. B., et al., in *12th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2004*, IEEE, Piscataway, 2004. With permission.) (©2004 IEEE.)

the plate and the wafer experience similar temperature cycles, which means that the device pattern faces a surface at almost the same temperature. As a result, the pattern effect is suppressed. This approach is practical for all conventional RTP processes except for fast spike-anneals that require ramp rates above 150 K/s [178].

11.3 Semiconductor Processing Using RTP

Over the last decade, RTP has progressed from being a technology that was used only in niche applications to being the dominant thermal processing technology in IC fabrication [179,180]. This change has come about for a combination of reasons. Firstly, as device technology progresses, there is an ever-increasing need to limit the diffusion of dopant distributions during thermal processing steps. This leads to a continuous reduction in the thermal budget that can be employed in device manufacturing [181]. However, the need to minimize defects and to form high quality films often demands that high temperatures are employed. Rapid thermal processing approaches can often overcome this dilemma by enabling the application of high temperatures for short times. This point is most obvious in the formation of the shallow junctions needed for source/drain extension regions, but also arises in silicide processing and other applications. The second factor behind the rise of RTP has been the ever-increasing need for improvement in process control and reduction of defects. Continuous advances in RTP temperature control capability have allowed the range of applications of RTP to expand from the early days of niche applications such as ion-implant monitoring to the most critical aspects of the device fabrication process such as gate dielectric processing and ultra-shallow junction (USJ) formation. The ability for rapid exchange of process gases and for excellent control of gas composition and purity has also been critical for many RTP processes [182–184]. This aspect motivated the early adoption of RTP for silicide formation, and has also been recognized as being essential for the formation of high-quality dielectrics and for shallow junction formation. The combination of excellent thermal uniformity and gas ambient control means that RTP systems offer significant improvements over conventional furnaces. This trend is greatly reinforced by the adoption of large diameter wafers, since the "planar" heating geometry in RTP tools is inherently easier to scale to larger sizes that the "tubular" form of conventional furnaces, which heat the wafers from the edges. The third factor driving a rapid

expansion of RTP is the great advantage in cycle-time that comes from a single-wafer heating strategy relative to conventional batch furnaces [180]. Several studies have demonstrated the economic benefits of replacing batch furnaces with RTP, especially in foundry operations, where a wide mix of products must be handled and batch sizes may be relatively small [185,186]. The fast cycle time of RTP is also useful for rapid development of new manufacturing processes, which is especially important when new materials are being integrated.

Another aspect where RTP differs from conventional processing is in the nature of the heating itself. For lamp-based RTP, the wafer is typically exposed to a flux of relatively high energy photons, as compared to the situation in a hot-wall system. Since the early days of RTP there has been some controversy about whether "photonic effects" influenced the outcome of the process [187,188]. Differences between process results obtained in various different RTP systems have sometimes been suggested to arise from photonic effects, although sometimes the effects of errors in temperature measurement, differences in the shapes of temperature-time cycles, and the presence of impurities in the gas ambient may be more important. Nevertheless, there certainly seems to be a capability for optimizing processes through consideration of the illumination conditions, and the progressive move to lower processing temperatures will make these issues more significant [189–191].

Tables 11.1(A) and Table 11.1(B) summarize the progress of RTP applications in both silicon device fabrication and in non-silicon applications, respectively [179]. Figure 11.52 illustrates the temperature–time domain of various key applications of RTP, and shows how the applications are evolving, mainly driven by the need for reduction of thermal budget [179]. A full review of the use of RTP in device manufacturing is beyond the scope of this chapter, and the interested reader is referred to several excellent

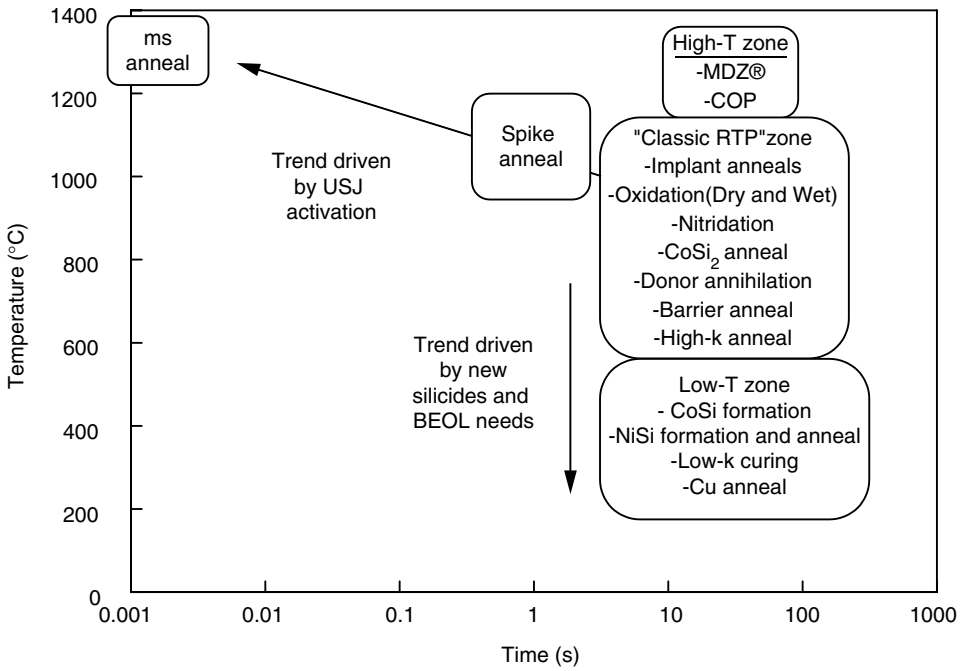


FIGURE 11.52 The temperature–time domain of RTP continues to evolve. Recent trends include "hotter and faster" processes that enable the creation of advanced ultra-shallow junctions (USJ) and low-temperature processing that is needed for NiSi formation and annealing. Wafer engineering processes, such as Magic-Denuded Zone® (MDZ, Magic-Denuded Zone is a registered trademark of Monsanto Electronic Materials Company) formation and crystal-originated pit (COP) annealing call for very high RTP process temperatures. (Reprinted From MacKnight, R. B., et al., in *12th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2004*, IEEE, Piscataway, 2004. With permission.) (©2004 IEEE.)

texts that cover the field in detail [1,2,192]. Instead this section will cover the motivation behind the use of the key RTP applications and recent innovations in the main process applications.

11.3.1 RTP Applications in Dielectric Formation and Processing

One of the earliest applications of RTP was in the formation of thin dielectric films, for example through rapid thermal oxidation (RTO) [193–195]. Rapid thermal oxidation approaches have been used for a very wide variety of applications, as indicated in Table 11.1(A). The greatest focus has been on processing of critical films, such as the transistor gate dielectric and capacitor dielectrics, but there are many other places where RTO presents advantages, such as in formation of shallow-trench isolation (STI) liner oxides, sacrificial oxides and in reoxidation of gate structures after gate etching [196–198].

Rapid thermal oxidation allows the use of relatively high process temperatures for the growth of silicon dioxide films while constraining the film thickness, an approach that is not viable in conventional furnaces. The high process temperature can lead to better quality films, and it has been suggested that the benefit may, in part, arise from the lowered film stress that can be achieved by growing the film at temperatures where the oxide film can flow [199–201]. Rapid thermal oxidation has also been shown to provide better film topography on trench structures, with reduced thickness non-uniformity at the corners of the trench, and with reduced crystal-orientation dependence [202–204]. For example, the formation of STI structures requires careful engineering of the trench geometry, especially with respect to the corner regions where stresses can generate defects in the silicon substrate. In these regions, the trench liner oxide thickness also tends to vary as a consequence of the impact of stress and crystal orientation on the oxidation rate. Rapid thermal processing allows the use of high oxidation temperatures for short times, which leads to more uniform films and also allows less time for stress-induced defects to evolve in the silicon. In recent years there has been great interest in RTO performed in a gas ambient containing steam, as a means for providing improved film characteristics, modifying growth rates and enabling selective oxidation applications [197].

11.3.1.1 Kinetics of Rapid Thermal Oxidation (RTO) and Nitridation (RTN)

Despite many years of research, the mechanisms of oxidation in thin oxide films remain controversial. Many studies have reported kinetics for RTO in various regimes of temperature, time, pressure, and ambient composition [147,193,195,205–213]. Early studies of RTO kinetics may be compromised by temperature measurement errors, which used to be a significant problem for RTP systems [147,207]. Many other factors can also affect the growth rate and the oxide quality in RTO processes, including the list of factors mentioned in Table 11.17. Certainly, aspects relating to surface preparation become very significant in oxides <5 nm thick. For such thin oxides, the sensitivity to temperature is actually less than for relatively thick oxides, and process uniformity limits are usually set by wafer quality and surface preparation.

Figure 11.53 shows curves illustrating the growth of oxide at various temperatures and times in oxygen at atmospheric pressure. Figure 11.54 shows the growth rate in O_2 as a function of the inverse of the temperature, and suggests that the activation energy is $\sim$2.1 eV for these films, which were $\sim$10 nm thick [195]. A study of spike-RTO with thinner films, in the range from 1.5 to 4 nm suggested an activation energy of 2.5 eV [206]. Figure 11.55 shows some growth curves for oxidation at reduced pressure [212]. It is important to note the behavior at low partial pressure and high temperature, where there can be a transition from passive oxidation, which leads to the growth of a SiO_2 film, to active oxidation, which leads to formation of SiO [205,214,215]. In active oxidation, the volatile SiO evaporates from the wafer surface and thermal etching occurs. This etching can be observed as haze on wafer surfaces. The phenomenon is important to understand since it illustrates the significance of ambient purity in many high temperature processes. Small concentrations of oxygen or water vapor, which are often present in "pure" process gases or from out-gassing of chamber components, can have a profound influence on many high temperature processes, including applications where oxidation is not the prime objective. Very thin oxides, such as native oxide, also tend to be unstable at high temperature as a result of

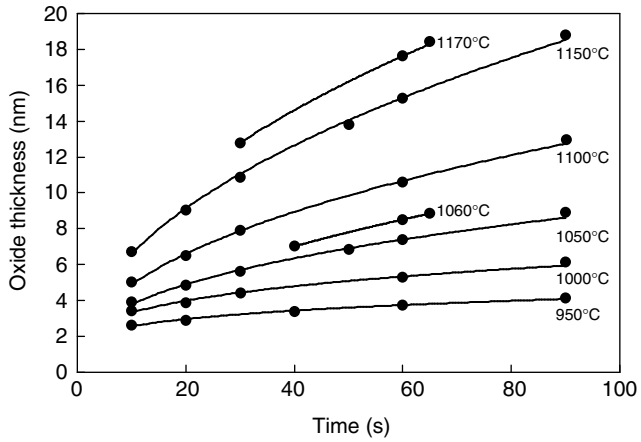


FIGURE 11.53 Rapid thermal oxidation (RTO) of (100) silicon in dry O_2. The solid lines are guides for the eye.

these phenomena [216]. The deliberate control of ambient, for example, by introducing several hundred parts per million of O_2, can be very useful in eliminating thermal etching [184].

There have also been extensive investigations of the kinetics of oxidation in ambients containing steam [217–219]. Figure 11.56 shows curves illustrating the growth of oxide at various temperatures in a 90% steam ambient [217]. Steam greatly increases the growth rate, and hence can be useful in lowering the thermal budget for oxidation.

Various other oxidation ambients have also been explored, especially those where the oxidizing species include gases containing nitrogen, such as N_2O and NO, and mixtures of these gases with O_2. Figure 11.54 shows a comparison of the oxidation rate in O_2 and in N_2O [195]. For temperatures

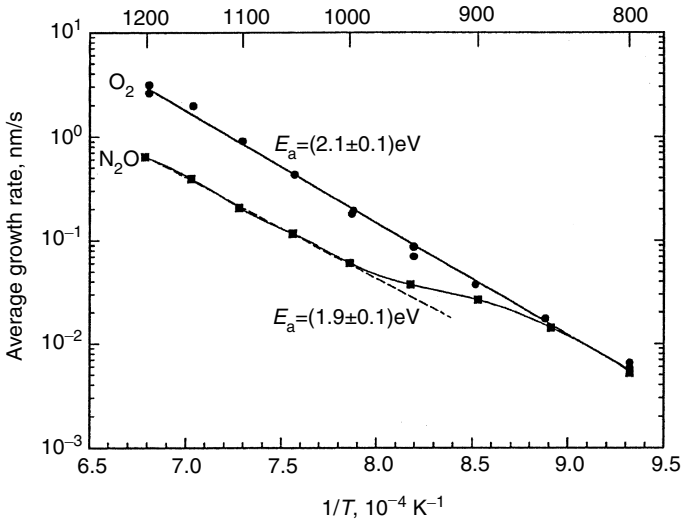


FIGURE 11.54 The kinetics of silicon oxidation in O_2 and N_2O for oxides that are ~10 nm thick. (Reprinted From Figure 12 of Green, M. L., in *Advances in Rapid Thermal and Integrated Processing*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996, 193. With permission.)

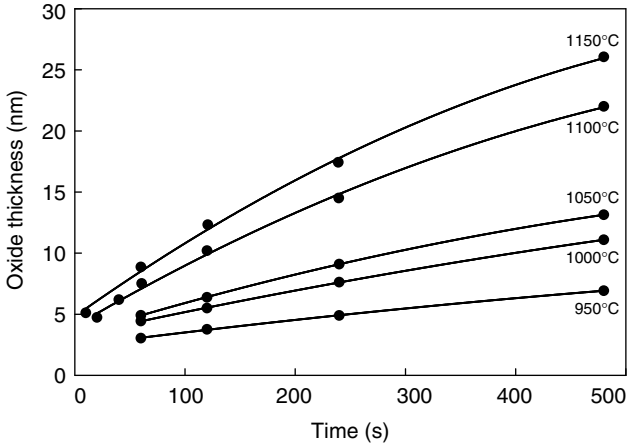


FIGURE 11.55 Rapid thermal oxidation of silicon in O_2 at 0.1 atm. pressure. (From Lassig, S. E. and Crowley, J. L., *Mater. Res. Soc. Symp. Proc.*, 146, 307, 1989.)

between 1000 and 1200°C, the N_2O oxidation rate is about one fifth of that for O_2. This is thought to be a consequence of the incorporation of nitrogen at the silicon/oxide interface, which will be discussed further below. The oxidation rates in both gases become very similar at temperatures below ~850°C, probably because nitrogen is not incorporated below this temperature. However, studies at lower

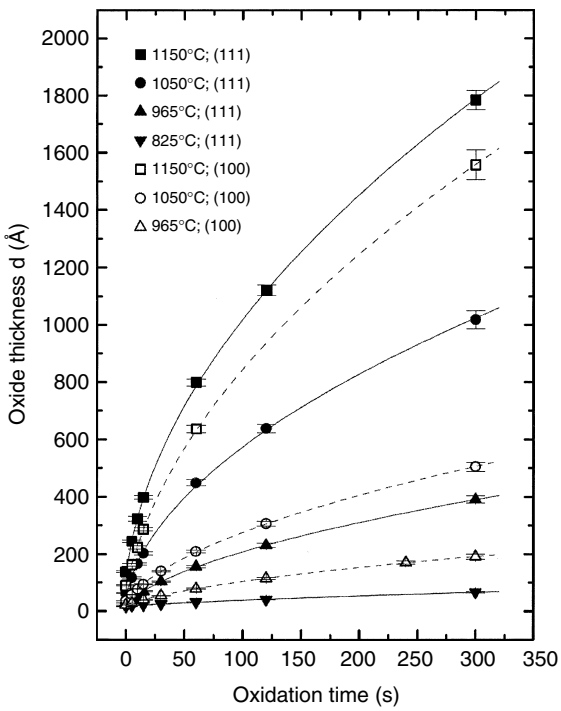


FIGURE 11.56 Rapid thermal oxidation in a 90% steam ambient. (Reprinted From Lerch, W., Roters, G., Munzinger, P., Mader, R., and Ostermeir, R., *Mater. Sci. Eng.*, B54, 153–160, Copyright 1998. With permission.)

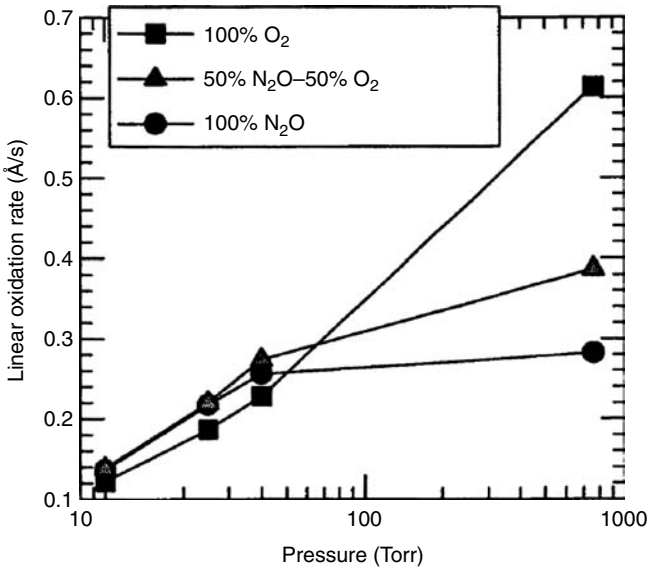


FIGURE 11.57 The pressure dependence of oxidation in O_2 and N_2O. At low pressure the N_2O oxidation rate can exceed that for an O_2 ambient. (Reprinted From Grant, J. M. and Karim, Z., *Mater. Res. Soc. Symp. Proc.*, 429, 257, 1996. With permission.)

oxidation pressures show that the behavior is more complex, and the N_2O oxidation rate can exceed that for O_2, as shown in Figure 11.57 [221]. This effect is believed to be a symptom of the complex gas-phase chemistry in N_2O oxidation, where N_2O breaks down to form atomic oxygen and NO species. The interplay of these species with the oxidation process, and the temperature and time evolution of their concentration in the reactor affect both the oxidation rate and the incorporation of nitrogen in the oxide film [221]. Studies of oxidation in NO have also been performed. In this case the oxidation rate is very slow, probably because a large amount of nitrogen is incorporated in the oxide, which then acts as a diffusion barrier to oxidizing species. The effect leads to a self-limiting growth [226,227]. Rapid thermal oxidation has also been performed in mixtures of NO and O_2, and a model was developed to explain the trend [228]. This method gives growth rates that are much faster than for NO oxidation without O_2. The effect of NO-annealing on the thickness of pre-existing RTO film has also been studied [226,230]. Typical NO-processes lead to a small increase in oxide thickness. The kinetics of RTO have also been reported for oxidation in combination with a variety of other gases, including halogen-bearing gases [193,231].

Silicon that is exposed to nitrogen-bearing gas will typically grow an oxynitride film, because the thermodynamics of the silicon–oxygen–nitrogen system make it extremely difficult to form a pure nitride layer when there is even the tiniest amount of oxygen-bearing gas present [232]. However, the reaction of silicon with ultra-pure ammonia gas results in a film that is mainly silicon nitride [233,234]. At any given temperature, the nitridation tends to be self limiting, because silicon nitride is an excellent diffusion barrier. Figure 11.58 shows an example of RTN kinetics for very thin nitride films formed in ammonia [234]. Silicon has also been nitrided by RTP in N_2 or N_2/O_2 mixtures [235–237]. For processing in "pure" N_2, the behavior is complex, because the formation of silicon nitride or oxynitride films may be governed by the presence of extremely small concentrations of background impurities in the gas ambient, especially H_2O, CO_2, and O_2, which may be present as a result of outgassing from chamber walls [235]. Several studies have reported the growth on SiON films in N_2 ambients, and this approach provides a way to form very thin oxynitrides with relatively high N content.

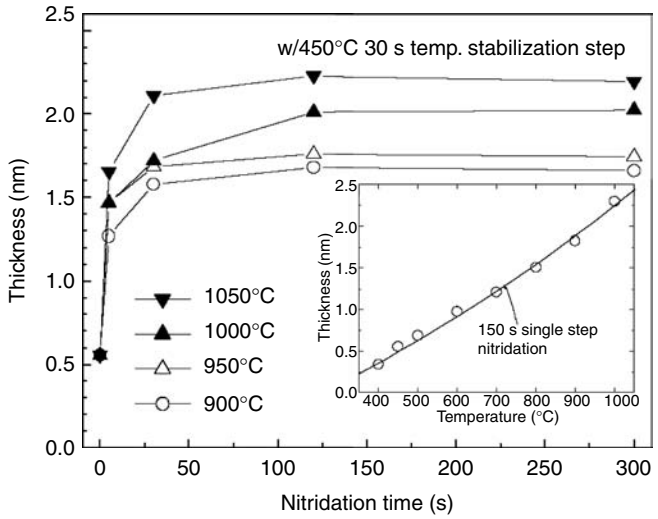


FIGURE 11.58 Rapid thermal nitridation of silicon in ammonia at atmospheric pressure. (Reprinted From Buchheit, K. M., Takeuchi, H., and King, T.-J., *Mater. Res. Soc. Symp. Proc.*, 786, E2.2.1, 2004. With permission.)

11.3.1.2 The Role of RTP in Forming High Quality Thin Dielectrics

11.3.1.2.1 Technological Benefits of RTP for Forming Thin Dielectrics

Advanced semiconductor devices rely on the performance and quality of extremely thin dielectric films, especially the gate dielectric in the metal oxide semiconductor (MOS) transistor [238,239]. Other critical films include capacitor dielectrics in dynamic random access memory (DRAM) as well as tunnel and interpoly oxides in flash memory devices. Rapid thermal processing has shown itself to offer many technical advantages in forming these dielectric films, and many papers have covered electrical properties and device applications [193–204,240–252]. As device scaling has progressed, the thickness of the films has reduced to the point where gate dielectrics are only ~1 nm thick [250–252].

The use of RTO to grow films at relatively high temperatures presents advantages in terms of smoother films, lower interface stress, reduced leakage, reduced trapped charge and interface state density, improved carrier mobility, and reduced hot-carrier degradation [199–201,240–243]. Furthermore, the reduction in thermal budget has become increasingly important as devices have scaled down in size to the point where diffusion of channel implants during gate oxidation is a significant concern. Reduction in thermal budget is also a key issue for many strain-engineering approaches and for incorporation of new materials such as SiGe and Ge [252–256].

A wide variety of RTP approaches have been evaluated for dielectric formation and annealing, including dry oxidation, steam oxidation, oxidation in ambients that introduce nitrogen and oxidation in ambients containing halogen-bearing gases [193–198,217–237,242–286]. Because of the small chamber volume in most RTP equipment, gas composition can be rapidly switched to allow sequential processing in different gas ambients [248,273]. Thermal profile design also offers new possibilities, such as spike oxidation processes [206,272,273]. The special benefits of forming dielectrics in steam ambients will be discussed in more detail in the section on steam-RTP below.

11.3.1.2.2 Improvement of Oxides

Rapid thermal processing approaches have also been used extensively in order to improve the quality of oxides [193–195,220–237,242–252,271–286]. The oxides may have been formed by RTO or by furnace oxidation processes. The RTP treatments have included relatively simple annealing processes, but greater benefits have been found from anneals in various process gases, most typically as means for introducing

nitrogen into the oxide. In such "nitridation" processes, the gases employed have included NH_3, N_2O, and NO [232,276–286]. As discussed above, nitrided oxides can also be formed by direct reaction between silicon and these gases. The nitrogen improves the oxide films in several ways. For example, introducing a few atomic percent of N into the oxide improves its ability to prevent the diffusion of B from the heavily B-doped gate electrode in a *p*-channel metal-oxide semiconductor (PMOS) device through the oxide and into the channel [232,282]. This B penetration can cause threshold voltage shifts and also degrades dielectric reliability. Relatively low concentrations (0.4–1 atomic %) of nitrogen can also improve some electrical qualities such as hot-carrier immunity [280]. As device scaling has progressed increases in nitrogen content have also allowed increases in the dielectric constant, which enable the use of thinner oxides while minimizing the leakage current from tunneling [238].

Optimization of the nitrogen content is a delicate task, because excessive concentration of nitrogen near the Si/SiO_2 interface can degrade the interface quality leading to increased interface states and mobility degradation [232,238]. As a result, control of the nitrogen distribution within the film is important. Such problems may be alleviated by very limited reoxidation of nitrided films, which regrows the interface region. Reoxidation or high temperature annealing steps are also useful after anneals in NH_3, because they help to eliminate H from the film [232,284–286]. The H is troublesome, because it can degrade hot-carrier reliability performance. Reoxidation moves the peak nitrogen away from the dielectric–silicon interface and minimizes the influence of interfacial nitrogen on carrier mobility. It can be performed in a dry or a wet ambient. The oxide grown in the reoxidation step determines the interface properties of the reoxidized dielectric stack. Therefore, the use of high-temperature, short-time RTP reoxidation provides a better quality interface [284,285].

The problems with H introduced by NH_3 anneals led to strong interest in the use of N_2O and NO for nitridation treatments, both for direct reaction with silicon, and for nitriding oxides [220–230,232, 233,242,245–248,271,272,280–283]. Since the amount of nitrogen incorporation with these gases increases rapidly with growth temperature [226], adequate concentrations of nitrogen can be

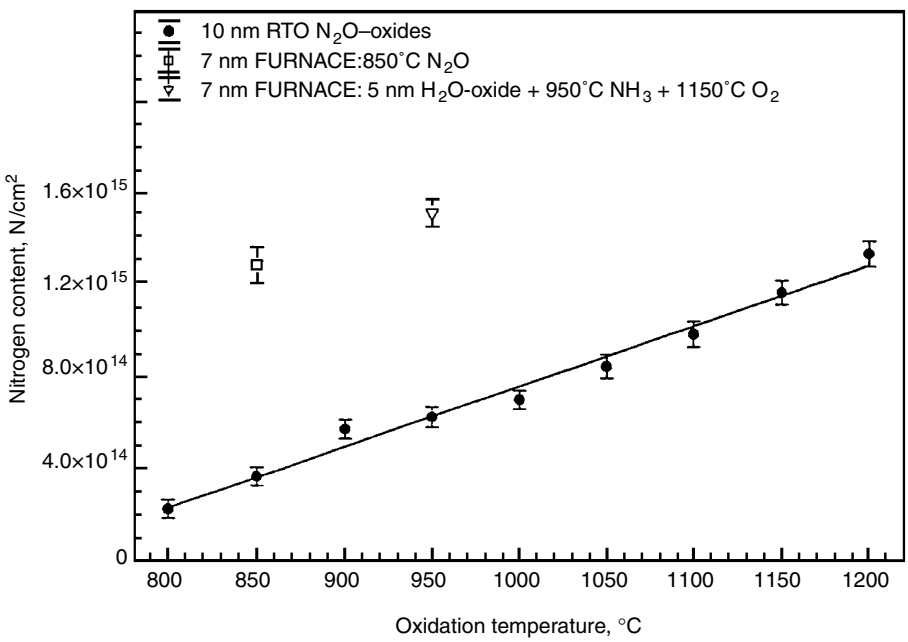


FIGURE 11.59 The temperature dependence of nitrogen incorporation for oxides formed by RTP in N_2O. (Reprinted From Figure 7 of Green, M. L., in *Advances in Rapid Thermal and Integrated Processing*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 193, 1996. With permission.)

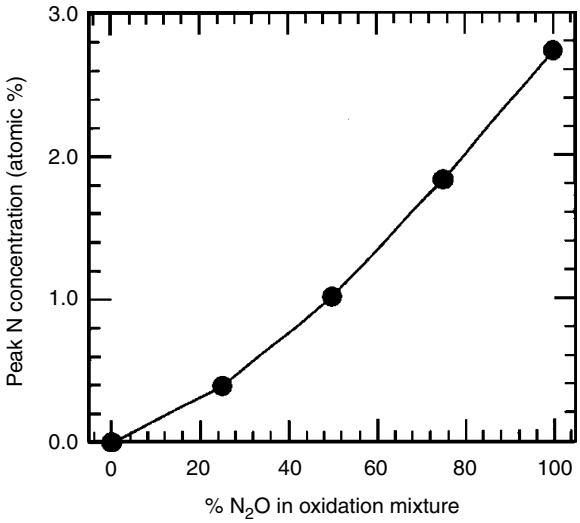


FIGURE 11.60 Peak nitrogen concentration at the oxide/substrate interface for 5.5 nm oxides grown in atmospheric pressure N_2O at 1100°C. (Reprinted From Grant, J. M. and Karim, Z., *Mater. Res. Soc. Symp. Proc.*, 429, 257, 1996. With permission.)

conveniently incorporated using high temperature process cycles achievable with RTP. Figure 11.59 illustrates the increase in nitrogen content with temperature for films formed in N_2O [195]. The degree of nitrogen incorporation also depends on the concentration of N_2O or NO in the gas mixture. Figure 11.60 illustrates the dependence of N incorporation in N_2O concentration for 5.5 nm thick oxides grown in N_2O at atmospheric pressure at 1100°C [221].

The quest for increased nitrogen content has led to adoption of new processing techniques including the introduction of nitrogen by plasma nitridation [287–290]. This approach allows a rather large concentration of nitrogen to be introduced, but it also requires a post-nitridation anneal in order to stabilize the nitrogen incorporation and improve film quality [160,287–290].

11.3.1.2.3 Rapid Thermal Nitridation of Silicon

An alternative approach for forming silicon nitride or oxynitride films with high nitrogen content is to form a very thin nitride film by nitriding the surface of the silicon in NH_3 [233,234,291–300]. The interface quality can be improved by reoxidizing the film, and good interface characteristics have been obtained by reoxidation in steam [291–293] or N_2O [294]. This approach has been shown to lead to highly uniform thin films with relatively large dielectric constants and very low leakage currents. Very thin nitrides formed by a sequence of nitridation and oxidation steps have recently been shown to be candidates for gate dielectrics in 22 nm node CMOS [296]. Thin nitride films formed by rapid thermal nitridation have also played an important part in forming capacitor structures, especially for DRAM applications. Often such films can be combined with CVD-deposited nitrides in a stack structure [297,298]. There has also been interest in using RTP for forming nitride films in silicon–oxide–nitride–oxide–silicon memory structures [249]. Very thin nitrides may also form useful interface layers for use in combination with high-κ dielectrics [299,300].

11.3.1.2.4 Role of RTO in Processing Trench Features

Rapid thermal oxidation has been shown to provide significant benefits in dealing with the challenges of forming uniform, high-quality dielectrics on trench features. For example, in the formation of trench capacitors for DRAM devices, stress concentration at the corners of the trenches can lead to oxide thinning, which can severely degrade the reliability of the device. Rapid thermal oxidation can be

performed at high temperatures, where the viscosity of the oxide is reduced and smooth and uniform oxide films can be formed even on trench features [202–204]. At the same time, the minimization of thermal budget helps restrain the formation and propagation of dislocations that might form in the stress regions. Rapid thermal oxidation methods have also proved to be very helpful for corner-rounding oxidations in STI structures [261,301]. Rapid thermal oxidation in steam ambients provides even more benefits for this application [197,198]. Rapid thermal oxidation can also be helpful in formation of sacrificial oxides [198,302].

11.3.1.3 RTP Integration with High-κ Dielectrics

Continued scaling of the gate dielectric is likely to lead to a need for a high-dielectric constant (high-κ) material with a dielectric constant that is even higher than that for silicon nitride ($\sim$7.5). In recent years a great deal of effort has been expended in the quest for suitable high-κ materials [238,239]. Currently the leading contenders are materials based on hafnium silicate ($HfSiO_x$) or hafnium oxide (HfO_2). In some instances, RTP may be used to form the high-k film by oxidation of a deposited film [303]. However, most research concentrates on the deposition of high-κ films by atomic-layer deposition (ALD) or metalorganic chemical vapor deposition (MOCVD). It seems likely that these materials will need to be deposited on an ultra-thin silicon oxide or oxynitride, in order to preserve the quality of the interface with silicon [239]. Very thin nitride layers may also be used at the interface [299,300]. Such "interface engineering" can benefit from the advantages of RTP in forming ultra-thin, high-quality films. Furthermore, deposited films typically need post-deposition anneals (PDA), in order to improve film stoichiometry by eliminating defects and impurities [304–310]. Such processes need to be performed with low thermal budget in a well-controlled gas ambient in order to prevent excessive oxidation, so RTP is an ideal approach. It also seems that nitridation of high-κ films will be important, in order to improve their thermal stability and diffusion barrier qualities. Nitridation can be performed by RTN in NH_3 or by plasma nitridation followed by post-nitridation annealing [311–316].

11.3.1.4 The Potential of Integrated Processing

Over the last decade, there has been great interest in the clustering of process modules, especially as a means for integrated processing of gate stack structures [317–324]. In principle, such approaches are expected to provide improved process control, for example, by minimizing the opportunity for contamination during transport between different processing tools in the factory. The single-wafer nature of most RTP approaches fits naturally with the cluster tool concept, and can be conveniently integrated with cleaning, deposition and plasma processing modules, as needed.

The ability to integrate cleaning is especially interesting for gate stack formation because the quality of thin oxides is critically affected by surface preparation [325]. Consequently, there have been many studies of the integration of cleaning steps with oxidation processes [318–322]. Cleaning methods have included in situ treatment with a variety of chemical species, and removal of organic contaminants and metals has been studied [319]. The removal of native oxides by vapor phase etching has been integrated into cluster tools [318–322]. In situ RTP treatment in reducing gases such as H_2 or GeH_4 has also received attention [326,327]. There has also been interest in clustering the gate dielectric formation step with a gate electrode deposition, for example, by RTCVD deposition of a polysilicon film. However, there has been less evidence that clustering this step gives a significant technical benefit, although there is data that suggests that clustering process modules can reduce defect densities [322–324]. The benefits of clustering have also extended to processes where a gate oxide is formed, followed by a plasma nitridation step and then by an RTP anneal that stabilizes the nitrided film [160,290].

Despite the great interest in such approaches, they have only recently become significant in the volume manufacturing environment. However, as the control of interface layers and minimization of contamination from the ambient becomes increasingly critical, and as advanced single-wafer deposition methods such as RTCVD and ALD are adopted, the trend towards clustered processing for critical films can be expected to continue.

11.3.1.5 Rapid Thermal Oxidation in Steam

Over the last decade, the introduction of steam ambients has further extended the range of RTO applications [196–198]. Steam oxidation enhances the growth rate for RTO, which may be useful for forming films where the process time would otherwise lie outside the range where RTP is economical. The enhanced growth rate also allows greater reduction of thermal budget. The presence of steam can also lower the viscosity of oxide films, leading to even lower stress and improved quality. Such benefits can also be useful in reflow of deposited oxides. The use of steam as an oxidizer also enables processes such as selective oxidation, which will be discussed in more detail below.

Steam can be generated and introduced to the RTP chamber by various methods. Early approaches included the use of bubblers, where process gas passes through a water cell before entering the chamber [259]. The steam concentration can be adjusted by changing the temperature of the water cell. Although useful for experiments, this approach does not provide very good control of the ambient conditions and is not used in manufacturing. Later approaches included the use of pyrogenic or catalytic steam generators [217,218,260–263]. Steam can also be generated by direct reaction between oxygen and hydrogen in the RTP chamber, stimulated by heat from the wafer [219,253,254,264,265]. Such in situ steam generation approaches have found a variety of applications, including gate dielectric formation, trench oxidation, and the oxidation of silicon nitride films [196,198,264]. Some properties of this style of oxidation may arise from the formation of highly reactive free oxygen radicals during the combustion process in the chamber [265]. Rapid thermal oxidation with steam generated from external steam generators has also found a wide range of applications, including thin dielectric formation, trench oxidation and selective oxidation [197,260–262,266–268]. Both oxygen and hydrogen-rich steam ambients have been explored. The hydrogen rich approach enables selective oxidation applications. The latter are of great interest for gate contact reoxidation in gate stacks with tungsten cladding for reduced gate resistance that are being evaluated for advanced DRAM technology. The selective oxidation approach enables the oxidation of the silicon to occur without oxidation of the tungsten.

11.3.1.5.1 Steam RTO Applications

It has been suggested that thin gate dielectrics formed by RTO with steam have improved electrical characteristics, for example, by exhibiting higher reliability and better breakdown characteristics [197]. It is believed that the improved qualities stem from the effect of the steam in reducing the viscosity of the oxide and reducing stress in the film. Steam RTO is also useful in applications where it is necessary to obtain good corner rounding on trenches, such as STI structures [196–198]. Once again, the improved ability for the oxide to flow, combined with the relatively high process temperature, can lead to better corner rounding, reduced film stress and reduced defects. Similar benefits have also been obtained in steam RTO for the formation of sacrificial oxides [198].

The in situ steam generation approach has also been shown to have some unusual capabilities that are believed to arise as a consequence of species generated by the reaction chemistry. One capability is the ability to oxidize nitride films [196]. Silicon nitride normally oxidizes very slowly in oxygen or in steam. However, it is believed that during the reaction between hydrogen and oxygen various more reactive species are formed and that these species are responsible for the accelerated oxidation rate.

11.3.1.5.2 Selective Oxidation

The ability to process wafers in an ambient containing both a reducing species, such as hydrogen, and an oxidizing species, such as steam, enables selective oxidation applications [266]. This is possible because steam does not react with hydrogen. The most important application to date is in the reoxidation of gate stacks that include a tungsten metal film [266–269]. In advanced DRAM devices, a tungsten strap is used over the polysilicon gate in order to reduce word-line resistance. After the gate etch, it is necessary to repair damage and to round the corner at the gate electrode/oxide interface. A selective oxidation process is required that does not oxidize the tungsten, yet does oxidize silicon to repair the damage from gate etching. Rapid thermal oxidation in a hydrogen-rich steam ambient enables this process while keeping a low thermal budget. The selective process is typically performed using 0.1%–10% steam in H_2 at

temperatures between 900 and 1050°C. Such processes may also find applications in processing of metal-insulator-metal capacitors [270].

11.3.2 Applications of RTP in Ion Implantation Damage Annealing and Dopant Activation

11.3.2.1 Rapid Thermal Annealing (RTA)

Rapid thermal processing originally evolved from intense interest in the late 1970s on the question of how best to anneal the damage introduced by doping through ion implantation. Early work on rapid annealing emphasized the potential of rather exotic annealing approaches that used a wide variety of energy beams, including lasers and electron beams, which were usually used to produce heating cycles with durations between nanoseconds and milliseconds [328,329]. Such approaches were some way ahead of the practical needs of device fabrication, and the "beam annealing" technologies remained technological curiosities for many years. However, it was also found that relatively short annealing cycles, with durations of a few seconds, at relatively high temperatures, $> \sim 950$°C, could anneal implant damage while resulting in significantly less diffusion than conventional furnace anneals. Such heating cycles could be conveniently obtained by heating the wafer with radiant energy from a bank of high power lamps. This style of RTP was also found to be very convenient for rapid evaluation of ion implanter performance. A wafer could be annealed and its sheet resistance measured in a very short time, possibly the first example of a great benefit arising from the cycle-time reduction inherent in RTP approaches [180].

As device technologies advanced, the need to minimize the degree of diffusion became more significant, and at the 0.25 μm device node, RTP became essential for implant annealing. Since that era, RTP has become an increasingly critical part of the processing needed to form the USJs needed in advanced CMOS devices [330–335]. The development of suitable RTP processes for implant annealing is a challenging task, partly because of the complexity of the materials science underlying diffusion and defect annealing phenomena, and partly because the annealing occurs relatively late in the transistor fabrication process, and hence it affects the properties of a wide range of pre-existing device structures. Such process integration issues can be very complex, and careful optimization is needed to obtain the best electrical characteristics. There are many ways of combining annealing cycles with the implants used for well and channel doping, source/drain extension doping, halo doping, "deep" source/drain contact doping and polysilicon gate activation [335]. Furthermore, the materials science issues for doping with *p*- and *n*-type dopants differ, which leads to further challenges in identifying an optimal sequence. Hence RTA processes may take many forms, depending on how the process integration is worked out.

One key advantage for RTA of implanted layers stems for the ability to reduce transient-enhanced diffusion (TED). Transient-enhanced diffusion describes the very large increase in dopant diffusivity observed in ion-implanted silicon, relative to the classical diffusivities for dopants in undamaged silicon [336–339]. This increase in diffusivity arises from the large excess of point defects that result from the ion implantation process. The phenomenon is especially severe for boron doping, since boron is already a fast diffuser, and its diffusivity is increased by silicon interstitials. Transient-enhanced diffusion effects are more severe at lower temperatures, because the degree of the excess of silicon interstitials over the equilibrium value (the super-saturation) is greater at lower temperatures than at higher temperatures. As a result, annealing at higher temperatures reduces the effects of TED, so long as the heating cycle can be kept sufficiently short. Rapid thermal processing enables the fast ramp-up and ramp-down in temperature that is essential for this approach [340–342]. Rapid thermal processing processes for USJ formation currently rely on spike-annealing, where the wafer is ramped to a high temperature at a fast rate, typically ~ 250 K/s, and then immediately allowed to cool [324–332,340–352]. Figure 11.61 shows secondary ion mass spectroscopy (SIMS) profiles for wafers implanted with $10^{15} BF_2^+/cm^2$ at an energy of 1.1 keV after state-of-the-art spike annealing [334]. Transient-enhanced diffusion effects may occur at several thermal processing steps in the process flow, but they can be limited by RTP annealing after

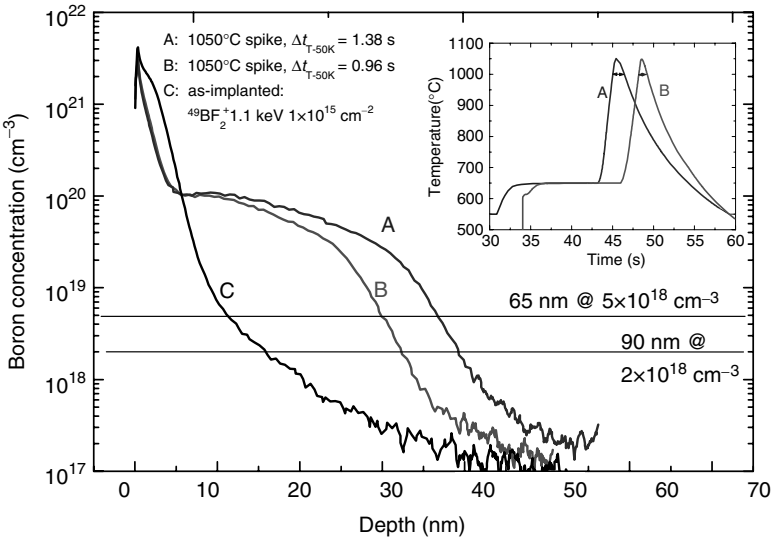


FIGURE 11.61 The impact of spike-anneal peak-width on diffusion is illustrated by secondary ion mass spectroscopy (SIMS) profiles from two 300 mm wafers that had been implanted with 10^{15} BF_2^+/cm^2 at 1.1 keV, and spike annealed at 1050°C (From Timans, P. J., Lerch, W., Paul, S., Niess, J., Huelsmann, T., and Schmid, P., *Solid State Technol.*, 47, 35, 2004). The time spent above 1000°C is 1.69 s (curves A) or 0.96 s (curves B).

implant steps, since this eliminates excess defects and hence prevents TED [353,354]. Many aspects of implant engineering can also be used to attempt to control enhanced diffusion effects, including the use of pre-amorphization implants (PAI) and co-doping schemes [333,339,355,356].

Rapid thermal annealing processes have been performed in a wide variety of gas ambients, and there can be strong interactions between reactions between the gas ambient and the silicon surface that can influence diffusion and annealing phenomena in the wafer [184,357–360]. The question of optimization of the annealing ambient has become increasingly important as implant energies have reduced and it is no longer practical to use screen oxides, since they absorb too much of the implanted dose. For example, Figure 11.62 illustrates the impact of oxidation-enhanced diffusion (OED) on implant profiles for wafers implanted with 10^{15} B/cm^2 at 1 keV and annealed for 10 s at 1050°C in N_2 ambients with varying oxygen content [184]. It is also important to realize that the oxidizing species that are inevitably present in the annealing ambient, even in the parts per million range, may also cause thermal etching. As a result, for B-implanted wafers it is often useful to anneal wafers in an ambient with a well-controlled oxygen concentration of a few hundred parts per million of O_2 [184]. This prevents thermal etching but does not enhance the dopant diffusion. Although, the effects of oxygen on diffusion can be rather large for soak anneals, the effect is somewhat less significant in spike annealing [349,351]. For some dopants there may be concerns about dose loss to the ambient during the anneal, in these cases a higher oxygen concentration may be used to even grow a thin oxide during the anneal process. For example, it is common to use between a few percent oxygen in the ambient when annealing As implants [357,360].

11.3.2.2 The Limits of Conventional RTP for Advanced Doping Technologies

The challenge in USJ formation is usually presented as the need for simultaneous minimization of both the junction depth X_J and the sheet resistance R_S. Although this X_J/R_S metric does not fully capture the complexity of the issues, for example, it does not directly address issues of lateral abruptness or defect annealing, the X_J/R_S trade-off does illustrate the central problem of the era: How to increase the concentration of electrically active dopant species without introducing excessive diffusion?

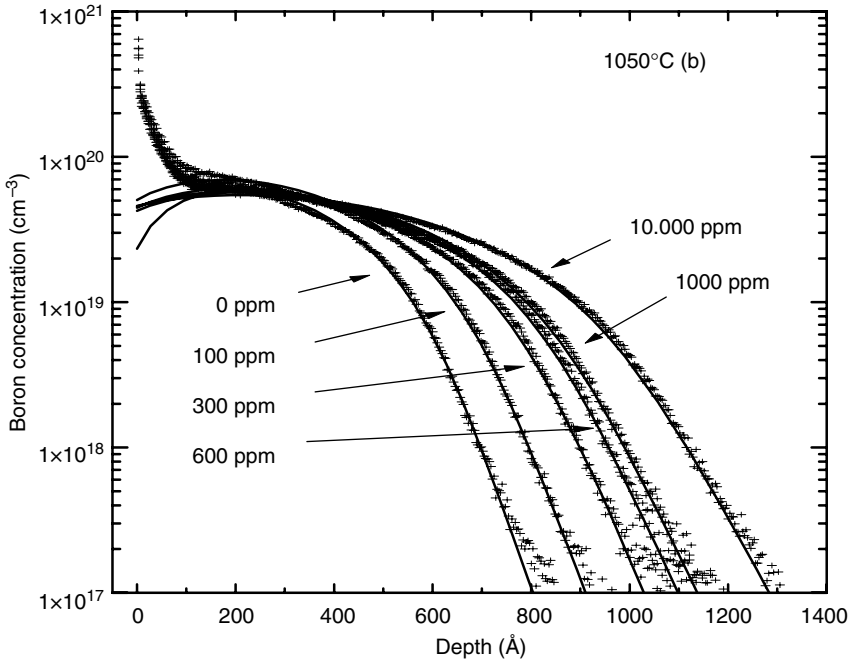


FIGURE 11.62 The effect of O_2 concentration on boron diffusion during 10-s soak anneals at 1050°C. The implants were 10^{15} B/cm^2 at 1 keV. (Reproduced From Lerch, W., et al., *J. Electrochem. Soc.*, 146, 2670, 1999. With permission.)

11.3.2.2.1 Spike Annealing and Peak-Width Reduction

As devices have scaled down, the general trend has been to adjust the RTP process by increasing the annealing temperature while simultaneously reducing the time at temperature, e.g., by moving from soak anneals to spike anneals [333]. This trend can be viewed as being driven by the difference between the thermal activation energy for dopant diffusion and that for electrical activation [345,346]. Figure 11.63 illustrates the origins of the trend. The activation energy for intrinsic B diffusion is 3.46 eV [346]. The dashed curve represents the time taken to activate 50% of the carriers available from an implant of 10^{15} B/cm^2 at 250 eV. The activation energy for the latter process was deduced to be ~4.3 eV [346]. The solid curves represent the times needed at any given process temperature to induce various degrees of dopant diffusion. The degree of diffusion was estimated as being $\sim[4D(T)t]^{1/2}$, where $D(T)$ is the intrinsic diffusion coefficient for B at the process temperature T, and t is the process time. In real device processing applications the diffusion length could be significantly larger, because of various mechanisms that can accelerate the diffusion, but these curves serve as a simple guide to the minimum diffusion expected. Since the process of electrical activation of dopants has higher activation energy than that for diffusion, it is kinetically favored at higher temperatures. Hence we can reach the 50% activation point with less diffusion by annealing for a shorter time at a higher temperature. Various researchers have found that this trend holds for activation of B species implanted in silicon [329,345,346]. Figure 11.63 suggests that annealing this implant with less than ~2 nm of diffusion would require an annealing cycle of ~1 ms duration at a temperature just below the melting point of silicon.

Recent developments in conventional RTP technology have included a strong emphasis on further reductions in the time at temperature, for example, by reducing the "peak width" of spike anneals, as reflected in the time spent within 50°C of the peak temperature [334]. Figure 11.61 illustrates, how a reduction in the spike-anneal peak width from 1.38 to 0.96 s led to ~7 nm less diffusion, measured at

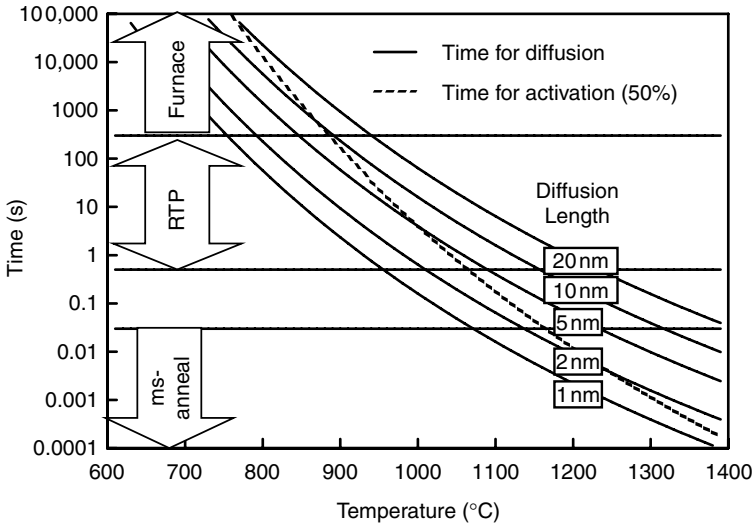


FIGURE 11.63 Illustration of the thermal budget criteria for various degrees of B diffusion and for electrical activation of 50% of the carriers available from 10^{15} B/cm^2 implanted at 250 eV. The diffusion lengths are minimal estimates based on intrinsic diffusion, and they do not include any enhancement effects. For advanced device technologies, where 1 nm of diffusion is significant, RTP processing is essential for most thermal steps. Activation of implanted dopants is only possible by very high temperature anneals with millisecond duration at temperatures just below the melting point of silicon. (Reprinted From MacKnight, R. B., Timans, P. J., Tay, S.-P., and Nenyei, Z., in *12th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2004*, edited by Gelpey, J., Lojek, B., Nenyei, Z., and Singh, R., IEEE, Piscataway, 3, 2004. With permission.) (©2004 IEEE.)

a concentration of 5×10^{18}/cm^3. However, at the same time the sheet resistance increased from 479 to 582 Ω/sq. If the peak temperature were to be adjusted to obtain the same target sheet resistance it was estimated that there would be only ∼1.5 nm of junction depth reduction. The peak width for conventional RTP systems is usually limited by the maximum cooling rate of the wafer and by the time taken to switch off the heating energy sources, which are usually tungsten–halogen lamps. These factors typically limit the peak-width to ∼1 s. More aggressive spike-width reduction is possible through the use of arc lamp energy sources, which can be switched off very fast and can provide spike anneals with peak widths of ∼0.3 s [347].

11.3.2.2.2 Millisecond Annealing

Such reductions in peak-width may be useful in the near term, but it has long been recognized that the trade-off between defect annealing and dopant diffusion illustrated in Figure 11.63 ultimately leads to a need for millisecond-duration heating cycles with peak temperatures somewhat below the melting point of silicon [329]. Some of the early "beam annealing" approaches mentioned above used CW laser beams or electron beams swept across the surface of the wafer to induce millisecond annealing. It was shown that adequate implant damage annealing and dopant activation could be achieved without introducing significant dopant diffusion, and that MOS devices could be fabricated using this approach [361]. The impending crisis in dopant activation has stimulated a renaissance of such "millisecond annealing" technology together with renewed focus on the challenges in materials science and in process integration. Most approaches are based on the use of pulses of radiation from banks of high-energy flash-lamps or from laser beams scanned over the wafer surface, although more exotic approaches, such as pulses of energy from high-power microwave sources, have also been considered [362–370]. The essential requirement is that the energy source only heats a region near the surface of the wafer during the duration

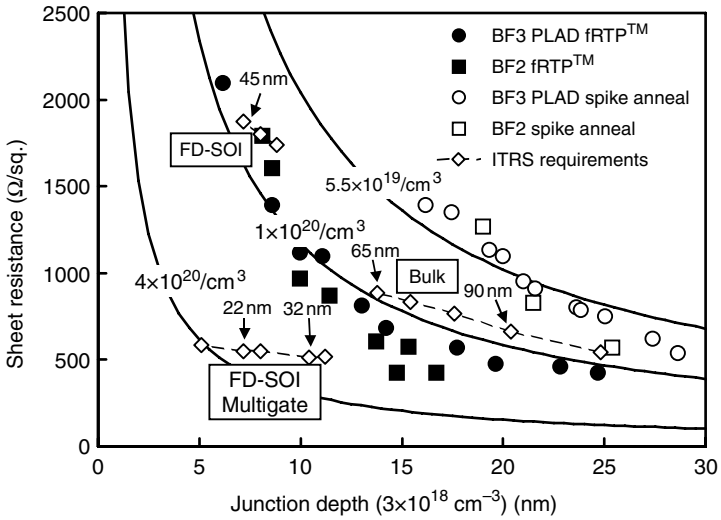


FIGURE 11.64 A comparison of the X_J/R_S capability of flash-assisted RTP™ (fRTP™) and conventional spike anneals against 2003 International technology roadmap for semiconductors (ITRS) specs. ITRS specifications vary for bulk, fully depleted SOI and multi-gate devices. Solid symbols are fRTP™ results for plasma-doping (PLAD) or beam-line implants. Open circles are the corresponding spike-annealing results. The solid curves are predictions for ideal, box-shaped doping profiles with various levels of electrically active B (concentrations are marked on the curves). (Reprinted From MacKnight, R. B., Timans, P. J., Tay, S.-P., and Nenyei, Z., *12th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2004*, edited by Gelpey, J., Lojek, B., Nenyei, Z., and Singh, R., IEEE, Piscataway, 3, 2004. With permission.) (©2004 IEEE.)

of the energy pulse. This condition allows extremely fast cooling through conduction of heat away from the surface into the bulk of the wafer. Although such approaches require a radical departure from conventional technology, they are attractive because they appear to offer a solution for forming advanced junctions that does not require extensive changes in the process integration scheme [366]. Figure 11.64 compares results achieved using conventional RTP spike annealing to those from millisecond annealing with the flash-assisted RTP™ (fRTP) approach [179,363]. This method combines a fast ramp to an intermediate temperature with a pulse of energy from an array of powerful water-wall flash-lamps that produces a temperature jump at the surface of the wafer [362–365]. Figure 11.65 illustrates the nature of

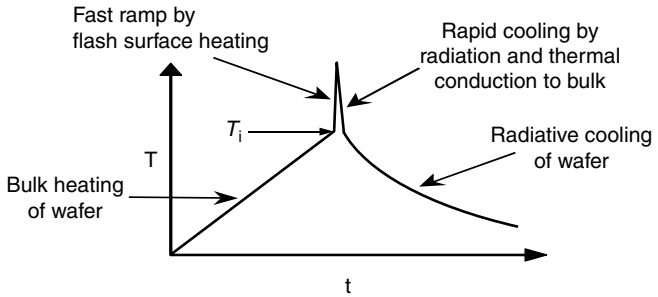


FIGURE 11.65 Millisecond annealing using the flash-assisted RTP™ method. In this approach, the wafer is rapidly heated to an intermediate temperature, T_i, and then a very short, high-energy pulse of flash-lamp radiation heats the whole front surface of the wafer.

the heating cycle. The study included a comparison of the results from samples doped by conventional beam-line implantation with BF_2 and samples implanted by a plasma doping approach with a BF_3 source gas. The fRTP results show a significant improvement in the X_J/R_S performance relative to conventional RTP spike annealing. The ability to adjust the intermediate temperature and the magnitude of the temperature jump induced by the pulse of lamp energy also provides flexibility in tuning the amount of diffusion relative to the amount of activation, which can help optimization of the very small degrees of diffusion that may be needed for tuning overlap with the channel. The figure also includes the X_J/R_S specifications for various types of CMOS device technology that are specified by the International Technology Roadmap for Semiconductors. It also includes a comparison of the process results with predictions of the X_J/R_S trends expected for box-shaped doping profiles with various concentrations of electrically active boron [343]. The concentrations for two of the curves were chosen so that the curves fell close to the X_J/R_S results from the spike anneals and the millisecond anneals. There is $>100\%$ improvement in electrical activation with the flash-assisted RTP approach. The curve shown for the concentration of 4×10^{20} B/cm^3 indicates the electrical activation needed to meet the needs of the 18 nm node. Further, refinements in millisecond annealing, such as the flash-assisted solid-phase epitaxy (SPE) approach, are expected to provide electrical activation at above this level [365].

Although the concept of millisecond annealing has been well-known for a long time, the challenges in moving it to volume manufacturing are significant. Quite apart from the process integration concerns, the usual issues of uniformity, repeatability, and wafer throughput pose new challenges. The issues in temperature measurement and control are also quite formidable, although some tools have demonstrated temperature measurement capability. The flash-lamp based tools have evolved from traditional RTP concepts and can process whole 300 mm wafers with a combination of a preheating step and a powerful flash of energy [363,366]. In contrast, the limited power and relatively poor efficiency in commercial laser systems requires an approach where a beam scans over part of the wafer and scan fields are "stitched" together [367]. This introduces fundamental concerns about short-range non-uniformity near the scan overlap regions. The finite thermal conductivity of silicon makes it impossible to create a temperature profile with very "sharp edges" so inevitably that some parts of the device die will be processed at different temperatures [369]. The effect can be somewhat reduced by operating with a very high degree of scan overlap, but this comes at the cost of a severe reduction wafer throughput.

Another challenge for all the millisecond annealing approaches comes from the pattern effect. Although this effect also exists in conventional RTP, the length scale for thermal diffusion in millisecond processing is far shorter, ~100 μm, meaning that temperature non-uniformities may arise from smaller features, such as scribe lanes, test areas and maybe even from different pattern densities within a die [332]. Several approaches can help reduce the effect. One method is to reduce the degree of variation in energy absorption in different regions of the wafer by heating with a relatively broad spectrum of energy, such as the output of a flash lamp. Another approach is to preheat the wafer and hence reduce the magnitude of the temperature pulse, and any associated non-uniformity. This approach also has the advantage of reducing the thermal stress introduced by the pulsed heating. The degree of preheating that is possible is limited by the amount of diffusion introduced by the preheating itself. The approach illustrated in Figure 11.65, which combines fast ramp heating immediately prior to the pulsed surface heating, and fast cooling immediately after the pulse, can greatly reduce the thermal exposure from preheating, while still allowing a relatively high preheat temperature [362–365]. It is also possible to form cap layers over the device structures to reduce or eliminate the pattern effect [371]. The disadvantage of the latter approach comes from added processing cost and complexity.

11.3.2.2.3 Pulsed Laser Annealing

Pulsed lasers, delivering energy over nanosecond time scales, have also been proposed for formation of very highly activated USJ structures [356]. In most studies the pulsed-laser approach requires the melting of a thin surface layer of the wafer that contains ion-implanted dopants. As the molten layer freezes back, crystal growth is seeded from the solid silicon and dopant atoms are incorporated on substitutional lattice sites at concentrations far above the solid-solubility limit. This approach was once viewed as being

very promising, but it has proved to be extremely difficult to incorporate it within a conventional process flow, mainly because the coupling of laser energy into device structure is very strongly dependent on the optical and thermal characteristics of the device structures [356]. As a result, device structures, especially in polysilicon gate and STI regions, are easily damaged by the very large stresses and melting phenomena induced by the high energy pulses. Many interesting schemes have been proposed to overcome these issues, and indeed the process window can be expanded through the use of absorber layers and energy control measures such as phase-change or reflectivity "switches" [372,373]. Unfortunately, these features introduce severe complications in the process flow, which may be difficult to accept in the manufacturing environment. Although the use of pulsed-lasers for solid-phase annealing remains an area of research, such laser systems usually operate in the nanosecond time-scale, which turns out to be somewhat too short to be useful for solid-phase processing [374]. Multiple-pulse solid-phase laser annealing can produce some of the benefits seen with millisecond annealing approaches, but it is not clear that such an approach could have a wafer throughput that is economic in manufacturing.

11.3.2.2.4 Solid-Phase Epitaxy

Another alternative for USJ formation is through exploiting the very high degree of electrical activation that occurs when dopants that have been implanted into an amorphous surface layer are incorporated in the lattice during epitaxial recrystallization of the amorphous layer [375–382]. Amorphous layers may be formed by the damage from doping ions or by a separate PAI. The PAI step is necessary for non-amorphizing dopants such as B, and it is often accomplished by implants of non-doping species such as silicon or germanium. In silicon, SPE is a thermally activated process that occurs rapidly at temperatures above $\sim$500°C [383]. The rate is also affected by many factors, including the type and concentration of dopants and impurities and by crystal orientation. Dopants are incorporated on lattice sites during the regrowth, and very high electrical activation can be achieved, with $\sim 3\times10^{20}$ carriers/cm^3 [379]. This is far above the equilibrium solid solubility at the SPE regrowth temperature, and the activation is metastable. Solid-phase epitaxy can be conveniently performed in conventional RTP tools, with process temperatures of $\sim$650°C and times $\sim$10 s [376]. The use of RTP is believed to allow some increase in the degree of electrical activation relative to the behavior at lower temperatures. There are several problems with process integration of an SPE approach, partly relating to the difficulties in obtaining simultaneous activation of polysilicon gates and in the complex interaction with the degree of activation of halo implants [375,381]. Another difficult issue arises from residual defects that remain after the SPE is complete. These defects are typically found in the end-of-range region, which lies just beyond the original amorphous/crystalline interface. They are not annealed during SPE, and if they lie in the depletion region of the junction they can cause a major increase in junction leakage. They can also cause problems during subsequent thermal processing, since they can act as sources of interstitials, which can induce further diffusion, or deactivation of dopants. These problems have limited the use of SPE in USJ formation, but there is still significant interest in the approach, because it allows the use of relatively low process temperatures that may be needed if new materials such as high-κ gate dielectrics and metal gates have to be integrated into CMOS [382].

Indeed it has been suggested that even greater electrical activation is possible by performing the SPE at very high temperatures [365]. The limited heating rate of conventional RTP limits the SPE temperatures to $<\sim$800°C, but millisecond heating approaches allow SPE at much higher temperatures. Studies suggest that this approach can produce extremely high activation for B implants, $\sim 6\times10^{20}$ carriers/cm^3 [365]. This approach may also present benefits in reducing the residual defects normally encountered with SPE.

11.3.2.3 Gate Doping

Rapid thermal annealing is also an important step in the formation of heavily doped polysilicon gates. Once again, the details of how RTA is used depend on the process integration scheme, but the important issue is maximizing the concentration of electrically active dopants in the gate. This aspect has become increasingly important in recent years, because the capacitance of the gate dielectric stack is reduced if the

polysilicon gate becomes depleted near the interface with the gate dielectric. Such depletion leads to a parasitic series capacitance, which lowers the overall capacitance, making the gate dielectric act as if it is thicker than its nominal value. Since further decreases of the gate dielectric thickness are very challenging, the performance loss from polysilicon depletion is very undesirable. Hence there has been a renewed focus on increasing the concentration of electrically active dopants in the polysilicon, and there is also strong interest in the use of poly-SiGe or even metal gates. The latter would be completely free from depletion effects because of their extremely high carrier concentration, but are far more difficult to integrate into manufacturing, as will be discussed further below. Until those complex issues are resolved, RTA processing can help in improvement of polysilicon doping, and once again high-temperature spike annealing has become a useful tool in obtaining the best electrical activation [340,384–387]. Furthermore, both millisecond annealing and pulsed laser annealing are being explored as means for improving polysilicon gate activation [388,389]. Typically, the millisecond annealing approach is combined with a spike anneal, because it is necessary to have significant dopant diffusion for the implanted dopant to have a chance to redistribute throughout the polysilicon gate and especially to reach the gate dielectric interface. Millisecond annealing has been found to improve gate activation, and the degree of carrier activation improvement has been shown to be equivalent to scaling the gate dielectric down in thickness by $\sim$0.2 nm. Pulsed laser approaches, where the film melts and dopants are incorporated during the rapid solidification, have also been demonstrated, although process integration may be difficult because of the issues discussed above.

11.3.3 Applications of RTP in Forming Contacts and Interconnect Structures

11.3.3.1 Silicide Formation and Annealing

For many years RTP has been used extensively in the formation of contacts for MOS devices. Although the benefits of RTP are usually thought of as arising from its enabling capability for thermal budget reduction, the ability to provide a very clean process gas ambient is also of great significance. Indeed, the first large-scale application of RTP for volume manufacturing arose in the formation of silicide films, a process that can be exceedingly sensitive to oxygen or water vapor contamination. Here, RTP systems, with their small chamber volumes, could easily provide very low oxygen concentrations that were problematic for large batch furnaces [390,391]. The RTP approach was also helpful because high annealing temperatures could be used, yet film agglomeration effects could be minimized [392,393]. Since that era, RTP has been the mainstay for formation of titanium silicide, which was followed by the widespread adoption of cobalt silicide at the 0.25 μm node, and now with the introduction of nickel silicide at the 90 nm node. These materials have typically been fabricated through the two-step self-aligned silicide (salicide) process [390]. The process starts with deposition of a metal layer, followed by an initial RTP process (RTP1) in which the metal reacts with silicon in order to form a silicide phase. The next step involves selective etching that removes unreacted metal and other by-products. The final step is a second RTP process (RTP2), which converts the silicide to the final phase of silicide that is desired. For both Co and Ni silicides, it was initially thought that a single-RTP step approach might suffice, but in practice the need to provide a robust manufacturing process led to adoption of two-stage approaches [394–397].

11.3.3.1.1 Titanium Silicide

Titanium silicide was the first silicide to be extensively used in the salicide process [390–393,398–408]. A two-stage RTP approach was adopted, where the C49 phase of $TiSi_2$ is formed in an initial low temperature anneal, and then this is converted to the low-resistivity C54 phase by a second, higher temperature, RTP step. This approach was adopted partly because the formation of the C54 phase requires a relatively high temperature anneal, and this condition could lead to reaction between the blanket Ti film and SiO_2 regions [391]. Another very important issue is that silicon is the species that diffuses during the formation of $TiSi_2$ and hence there is a danger that the silicide may form over the spacer, leading to "bridging", where conducting residue shorts the gate to the source/drain regions.

By using a two-step approach, the temperature of the first stage of silicide formation can be reduced, preventing the reaction of Ti with SiO_2 and restraining the diffusion of Si. Titanium is easily oxidized, so good control of the ambient purity is essential, and typically it is necessary to have $< \sim 5$ ppm O_2 in the gas ambient. Furthermore, the process is usually carried out in nitrogen, which leads to the simultaneous formation of a TiON surface layer [390,391]. This aspect is useful because the TiON can help to block rapid lateral diffusion of Si through the film and help prevent bridging. The RTP1 step is typically performed at temperatures between $\sim$650 and 720°C, for $\sim$30 s. Unreacted Ti metal and TiON is removed by a selective etch and then the RTP2 step is performed in order to convert the silicide to the C54 phase. The RTP2 step is typically performed at temperatures between $\sim$800 and 900°C, for $\sim$30 s. The RTP approach serves an essential role because the high temperature promotes rapid nucleation and growth of the C54 phase of $TiSi_2$ without introducing agglomeration [392–393]. Figure 11.66 illustrates the processes in the conversion from Ti to $TiSi_2$ through the trends of the sheet resistances of Ti films after RTP anneals over a range of temperatures.

The $TiSi_2$ approach faced several difficulties in extension to narrow linewidth features. In particular, as the linewidths and the silicide thickness reduce, it becomes increasingly difficult to convert the C49 phase to C54. The C54 phase nucleates at the intersections of grains of C49, and in very narrow lines the density of such nuclei becomes very small. In order to overcome this limitation the RTP2 temperature has to increase, but this leads to a greater chance of agglomeration, which destroys the integrity of the line. Hence, on very narrow features, the process window vanishes [393]. Many attempts were made to extend the use of $TiSi_2$ to narrower features, including approaches of including extra metallic elements such as Mo, PAI of the silicon and the use of high-ramp rate heating [404–408]. Despite these efforts, it was difficult to extend the technology below the 0.25 μm node, and $CoSi_2$ was adopted as the alternative solution.

11.3.3.2 Cobalt Silicide

Cobalt silicide was introduced at the 0.25 μm-node, in order to overcome the limitation of $TiSi_2$ in scaling to narrower feature sizes [393,408]. Three silicide phases are involved in the silicidation process, Co_2Si, CoSi, and $CoSi_2$. The mechanism of formation is quite complex, since both Co and Si species can diffuse during the process. Since silicon is the main species diffusing in the formation of the CoSi phase, the need to limit bridging led to the adoption of a two-stage RTP approach rather similar to that used for $TiSi_2$ [394]. In this case, the first RTP1 process leads to the formation of the Co_2Si and CoSi phases.

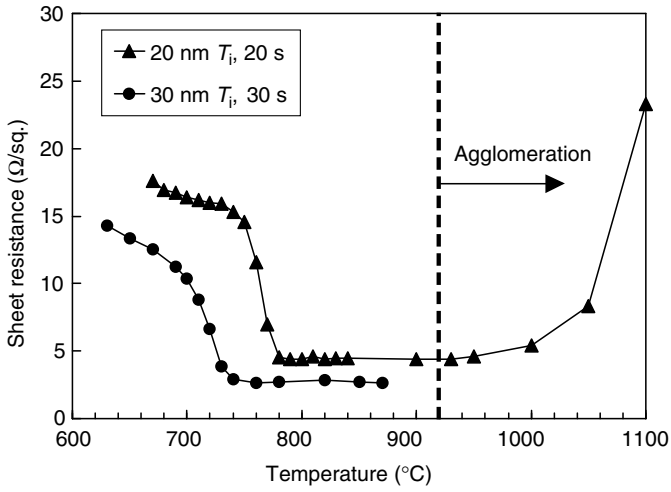


FIGURE 11.66 The effect of annealing temperature on the formation of titanium silicide.

The RTP1 step is typically performed at temperatures between ~420 and 550°C, for ~30 s. Co metal is prone to oxidation and excellent control of the ambient purity is essential. Indeed, early attempts to produce $CoSi_2$ in conventional furnaces occasionally resulted in the formation of cobalt oxide (Co_3O_4) instead [391]. Hence RTP approaches provide an important benefit in ambient purity, although it is also important to realize that even in a very pure gas ambient, gases can still desorb from the wafer itself and affect the process [409]. Various capping layers have been explored, especially Ti and TiN [409–416]. Ti caps have the benefit of being able to getter oxygen and the presence of Ti can also help to eliminate residual oxide on the silicon surface that can hinder the silicide formation [409]. However, the Ti interacts with the silicidation process itself and can lead to a need for a higher thermal budget for $CoSi_2$ formation [409,414]. The use of TiN caps has also been studied extensively and many device manufacturers adopted this approach [413–415]. After the RTP1 step, unreacted Co metal and the capping layer are removed by a selective etch and then the RTP2 step is performed in order to convert the silicide to the $CoSi_2$ phase, which has a lower resistivity. The RTP2 step is typically performed at temperatures between ~750 and 900°C, for ~30 s. Figure 11. 67 illustrate the trend in sheet resistance with annealing temperature for a 15-nm thick Co film with a 10-nm thick cap of TiN. Studies of the impact of RTP1 and RTP2 temperatures show a complex array of interacting physical phenomena at work, complicated further by the need for simultaneous optimization of the end result on n^+ and p^+ regions, single crystal and polysilicon. One common observation is that relatively high RTP2 temperatures reduce interface roughness and greatly reduce junction leakage currents [408,415].

The $CoSi_2$ approach faces limitations as linewidths reduce [395,396]. In particular, as junction depths decrease, the silicide thickness must be reduced, leading to increases in sheet resistance. This issue is quite severe for $CoSi_2$, on account of its relatively high silicon consumption. Furthermore, it is essential to obtain a smooth interface with the silicon. Interface roughness is determined by nucleation of the $CoSi_2$ phase, and roughness can be reduced by raising the RTP2 temperature, but on narrow lines this approach increases the sheet resistance [396]. As a result, the process window tends to vanish as linewidths scale below ~50 nm. Furthermore, $CoSi_2$ does not integrate well with SiGe materials, which have been adopted in the source-drain contact region at the 90 nm node and may also be used in the gate electrode [395,417]. The relatively high thermal budget needed for $CoSi_2$ formation also suggests that this approach may tend to deactivate dopants in the source/drain regions and gate electrode, which are increasingly important issues beyond the 90 nm node [415].

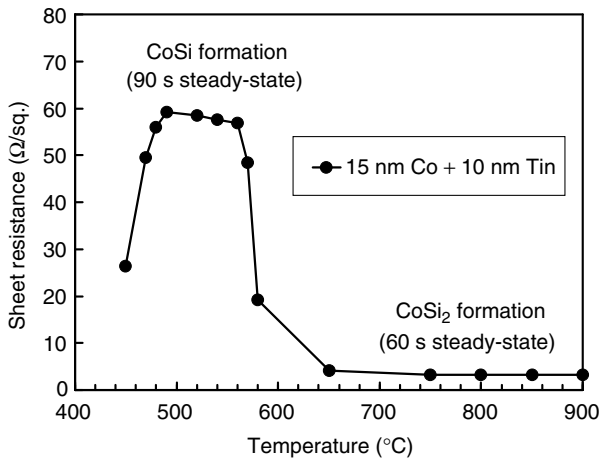


FIGURE 11.67 The effect of annealing temperature on the formation of cobalt silicide.

11.3.3.3 Nickel Silicide

Nickel silicide was introduced at the 90 nm-node, in order to overcome the limitation of $CoSi_2$ in scaling to narrower feature sizes, for improved compatibility with SiGe and for reduced silicon consumption [395–397]. The formation of NiSi poses severe challenges in materials science, because the Ni–Si phase diagram is very complex, with six phases that are stable at room temperature [418]. Furthermore, in situ microstructural studies of the behavior during temperature ramps reveal complex dynamics with a variety of phases appearing and transforming in a manner that depends on film thickness [419]. In NiSi formation Ni is the main diffusing species, but there is still some concern about bridging issues, and there is also a strong tendency for a "reverse-line-width" effect, where excessive silicide forms on small features which leads to a variety of problems, including junction leakage [395–397]. The origins of this phenomenon lie in the extremely rapid diffusion of Ni, which can lead to difficulties in process control. Once again, a two-step approach can overcome the challenge. In this case, the first RTP1 process leads to the formation of a nickel-rich Ni_2Si phase. This RTP1 step is typically performed at temperatures $\sim$300°C, for $\sim$30 s, although it has also been suggested that a low-temperature spike anneal at $\sim$360°C may be useful [396,420]. Some studies have shown that capping layers can also be useful in formation of NiSi films [421]. Unreacted Ni is removed by a selective etch and then the RTP2 step is performed to convert the nickel-rich silicide to the NiSi phase, which has a lower resistivity. The RTP2 step is typically performed at temperatures between $\sim$400 and 500°C, for $\sim$30 s. Figure 11. 68 shows an example of the full "conversion curve" for a 10-nm thick Ni film reacting with silicon as a function of RTP1 temperature for different processes in a two-step RTP annealing technology [422]. The soak time for the RTP1 step was 30 s. The ramp-up rate was set to 15 K/s and cooling rate was determined by the radiative cooling limit. The wafers were processed in N_2 and they were selectively etched after the RTP1 step. The sheet resistances of the wafers were measured before and after the RTP2 step of 30 s at 450°C. Wafers annealed at temperatures above 450°C during RTP1 were not subjected to the RTP2 step. At low temperatures Ni_2Si formation takes place during RTP1, indicated by an increasing sheet resistance compared to the as-sputtered Ni layer. With increasing temperature more of the Ni is converted to Ni_2Si, until at $\sim$280°C all of the Ni has been converted to Ni_2Si. This can be seen by comparing the post-etch curve with the post-RTP1 curve. Since the selective etch removes excess nickel, the etching increases the sheet resistance unless all the Ni has reacted with Si during the RTP1 step. In a small plateau between 280 and 300°C the

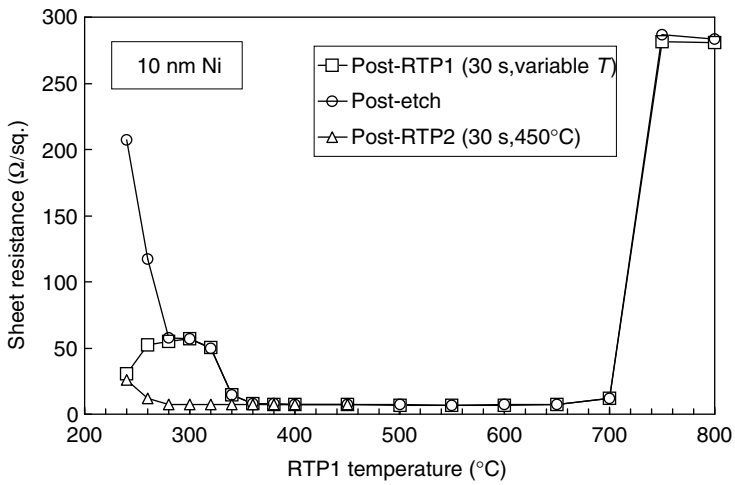


FIGURE 11.68 Temperature transformation curves for sheet resistance of a 10-nm nickel layer after the different steps of the process sequence of a two-step RTP technology. (From Niess, J., Paul, S., Buschbaum, S., Schmid, P., and Lerch, W., *Mater. Sci. Eng. B*, 114–5, 141, 2004.)

sheet resistance is independent of RTP1 temperature until the transformation of the Ni_2Si phase to NiSi starts at temperatures above 300°C, as indicated by decreasing sheet resistance after RTP1 and after etching. Above ~360°C there is an extremely broad range where the sheet resistance stays constant, until 650°C, where formation of the high-resistivity $NiSi_2$ phase starts. Transformation of the RTP1-annealed and selectively etched nickel silicide layers during RTP2 results in a constant low sheet resistance of ~7 Ω/sq. for RTP1 temperatures above 280°C. The final sheet resistance is independent of the sheet resistance after RTP1, provided that all nickel reacted during RTP1. For RTP1 annealing at about 280°C–300°C, the temperatures are low enough to avoid excess silicidation in narrow lines but at the same time high enough to provide a temperature insensitive process for uniform nickel silicide formation. The formation of $NiSi_2$ poses some special problems in this technology. In some circumstances this phase can also arise at low temperatures, even down to ~300°C, for example, as a result of epitaxy mediated by thin residual oxide at the Si/Ni interface [395–397]. The pyramid-shaped $NiSi_2$ grains are very undesirable because they increase silicon consumption and create a very rough interface [423]. Rapid thermal processing approaches that minimize time at high temperature may help reduce the possibility for $NiSi_2$ formation, but proper interface preparation is also essential [397,424].

NiSi technology also faces severe challenges in terms of thermal stability, especially with respect to agglomeration, and it requires that thermal exposure after silicide formation is minimized [395–397]. This is even more of a challenge for NiSiGe materials, whose thermal stability limits are typically ~100°C lower than those for NiSi. This makes the process window for NiSiGe formation smaller than that for NiSi [396,397]. Improved stability of both NiSi and NiSiGe may be achieved by including other elements, such as Pt, in the film [396,397,425]. Initially, the adoption of NiSi challenged RTP technology, on account of the need for careful process control at relatively low temperatures. However, modern RTP systems include advanced temperature measurement approaches that can measure wafer temperatures down to ~250°C, providing adequate margin for closed-loop control that is essential for repeatable processing [422].

11.3.3.4 Other Silicides

Rapid thermal processing has been used to form and anneal a very wide range of silicide materials, as described in several reviews [426]. The more technologically significant materials include WSi_2, which is often used in contact structures, especially in DRAM technology. In this application RTP is usually employed for annealing deposited films, rather than for forming the silicide [427–433]. The annealing process creates the correct microstructure and improves resistivity. New silicide materials may also emerge, for example, there is significant interest in the formation of $PtSi_x$ films as contacts for PMOS [434–436] and $ErSi_x$ [437] or $YbSi_x$ [438] films as contacts for n-channel metal-oxide semiconductor (NMOS) devices, since these combinations can lead to low potential barriers at the contact.

11.3.3.5 Titanium Nitride

Titanium nitride has long been employed in contact structures, where it serves as an excellent diffusion barrier and as a "glue" layer to enhance adhesion of tungsten plugs to oxide films [439]. One of the earliest RTP applications involved annealing deposited TiN films, and RTP can also be used to form TiN films, for example by the reaction of Ti with NH_3 or N_2 gas. As a result, there have been many studies of the formation and annealing of such TiN films [440–453]. Rapid thermal processing can also be used for grain-boundary "stuffing", where a diffusion barrier is annealed in a gas ambient that includes a reactive gas, typically oxygen, which is preferentially incorporated at the grain boundaries in the film. This tends to block the rapid diffusion paths along grain boundaries that could otherwise undermine the barrier properties.

11.3.3.6 RTP Applications in Interconnect Engineering

Rapid thermal processing has been applied for a variety of purposes in the back-end-of-line (BEOL) processing used to create multilevel interconnects. Although the application of RTP in the BEOL is in its infancy, applications may expand as the needs for low thermal budgets, and reduced manufacturing cycle-times become more significant.

11.3.3.6.1 Pre-Metal Dielectric Processing

One of the early applications of RTP was for reflow of deposited dielectrics such as boron–phosphorus–silicate glass (BPSG) [454–457]. Doped glasses are used as the first layer in the interlevel dielectric structure. Densification and reflow steps at elevated temperatures are performed after film deposition in order to achieve good planarization and void filling of the film. It is important to ensure that these steps have low thermal budget in order to minimize impact on device junctions and contact structures [456]. Rapid thermal processing can provide reflow properties equivalent to results from furnace annealing by using 100°C–200°C higher temperatures and much shorter process times ($<$60 s) [455]. The high-temperature, short-time step is desirable as it has a smaller impact on the underlying device layers than the longer, low-temperature step. Therefore, RTP can facilitate the pre-metal dielectric integration in a device manufacturing flow. Figure 11.69 shows the results of a systematic study of reflow angle as a function of RTP in comparison with furnace anneals [457].

Further reduction in thermal budget can be obtained by using steam in the reflow ambient [458–460]. The effect of steam on the flow angle is shown in Figure 11.70 [458]. 100% steam was used in the wet ambient. A lower flow angle corresponds to higher reflow. Similar results were reported in other studies of RTP reflow in a steam ambient [459,460]. The increase in reflow with steam is due to lowering of the glass transition temperature of BPSG. This decreases the viscosity of the BPSG film and promotes easier flow. The use of an oxidizing ambient (steam) increases the need for a short time process step to prevent oxidation of the subsurface layers [460]. This makes RTP ideal for BPSG reflow. Rapid thermal processing has also shown itself to be useful for the densification of deposited oxides used in STI structures [461,462].

11.3.3.6.2 Curing Low-κ Films

The continuing need to reduce the dielectric constant of the isolation layers used in multi-level metallization schemes has lead to the adoption of new low-κ materials. Such films often need to be stabilized by post-deposition treatment, which can drive out impurities and improve the adhesion and mechanical qualities. As a result there have been several studies of the use of RTP for curing low-dielectric constant films [463–465]. Rapid thermal processing approaches also present the advantage of reducing the degree of copper diffusion through diffusion barriers [463]. This may allow such barriers to be made thinner, which is desirable in order to minimize the interconnect resistance.

11.3.3.6.3 Copper Annealing and Reflow

Several studies have examined the use of RTP to anneal copper films [466–473]. Electroplated copper undergoes a self-annealing process at room temperature in which its resistivity drops to about 75% of its as-deposited value as a result of grain growth. This self-annealing process takes tens of hours to complete. The effect can be greatly accelerated by annealing the film at elevated temperatures. The effect of annealing conditions on sheet resistance is demonstrated in Figure 11.71 [466]. Samples annealed at 200°C for 1 s had the same sheet resistance as samples that were self-annealed for over 60 h at room temperature. Longer durations do not reduce the sheet resistance any further, and in fact could lead to diffusion of copper through adjacent layers. In addition, the oxygen levels in the annealing ambient have to be kept low to prevent oxidation of copper. These factors make RTP attractive for annealing copper films. The benefits of an RTP approach may also include improved grain structure and a reduced tendency for void formation. Rapid thermal processing has also been employed in novel copper reflow approaches [472,474].

11.3.4 Emerging Applications of RTP: Strain Engineering, Metal Gates, and Multi-Gate CMOS

As device scaling continues new applications for RTP continue to emerge. The need to introduce new materials and device architectures introduces new processing steps. Progress in scaling has continued through innovations such as the creation of strain in the channel, which can help increase the mobility of

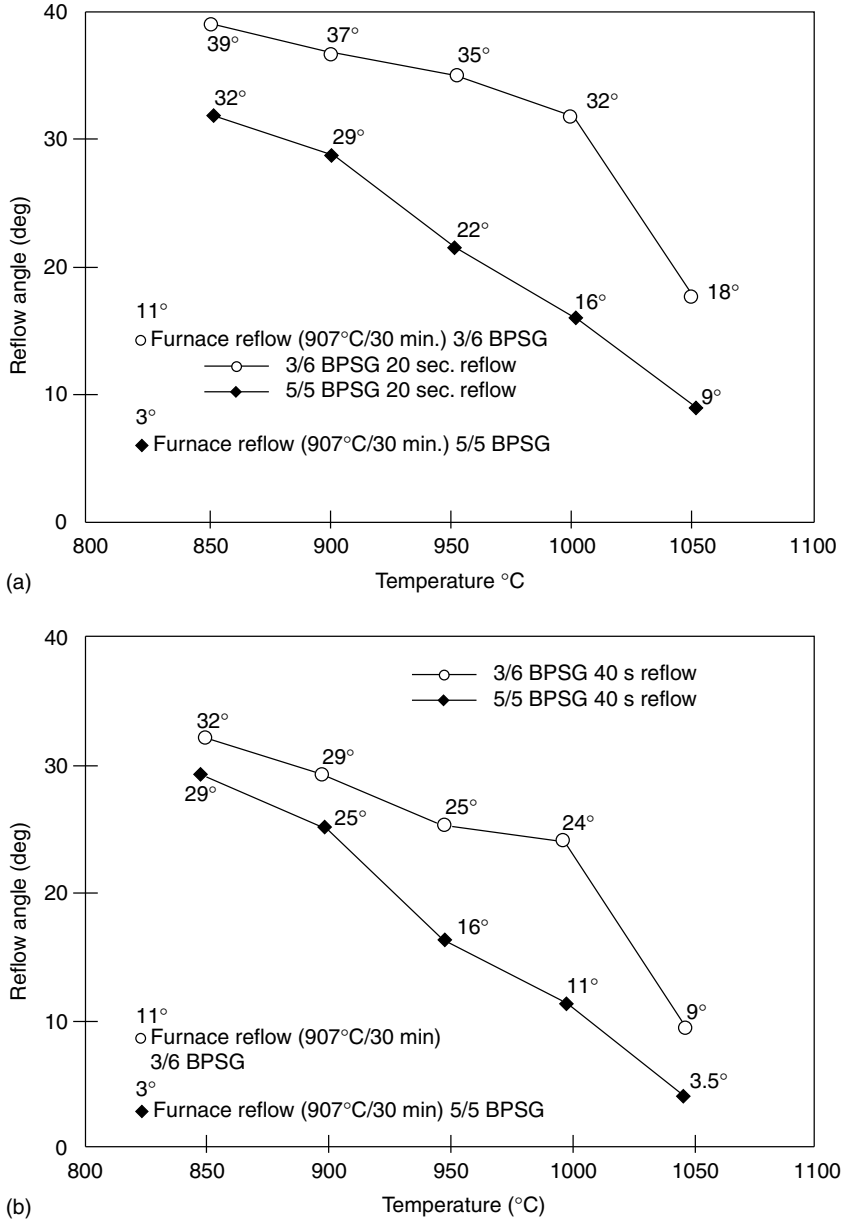


FIGURE 11.69 Variation of reflow angle for different RTP reflow conditions for 3B×6P and 5B×5P BPSG for (a) 20 s and (b) 40 s. (Reproduced From Iyer, R., et al., *J. Electrochem. Soc.*, 143, (1996): 3366, With permission of The Electrochemical Society, Inc.)

charge carriers and provide higher drive currents [474]. Strain engineering is now a key focus for advanced device technologies, and RTP plays a part in several ways. Rapid thermal processing approaches help to reduce the thermal budget, which may be helpful to prevent strain relaxation effects and to deal with issues in integrating SiGe films, which are often used in strain engineering approaches [253,254,475–479]. Rapid thermal processing can also be used to manipulate the state of stress in a wide variety of films [479–481].

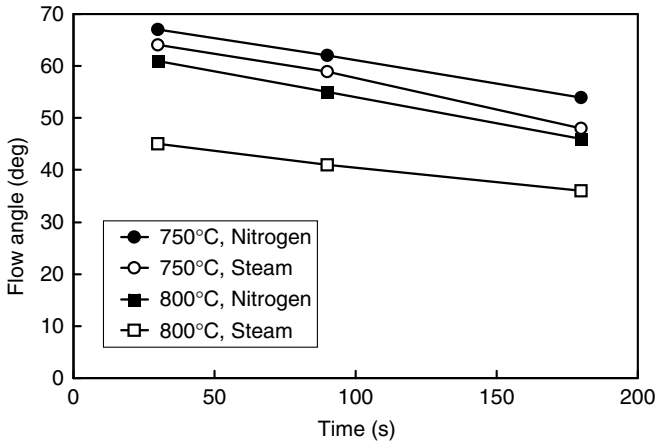


FIGURE 11.70 Dependence of flow angle of BPSG films on annealing ambient. (From Das, J.H., Brichko, Y., Daniel, A.D., Clarke, D., Kapkin, K., and Al-Lami, S., in *7th International Conference on Advanced Processing of Semiconductors—RTP 1999*, edited by Kitayama, H., Lojek, B., Miner, G., and Tillmann, A., RTP-Conference, 67 and 181, 1999.)

The move to metal gates also introduces new considerations for thermal processing. As mentioned above, the depletion effects in polysilicon gate electrodes can cause a significant reduction in the capacitance of the MOS gate stack in inversion. The depletion problem can be completely eliminated by replacing the polysilicon with a metal or silicide, where the carrier density is far higher and depletion effects are negligible. Metal gate structures may also be more compatible with new high-κ materials, which often form undesirable interface structures when contacted by polysilicon [482]. There may also be further circuit performance advantages from reduced resistance of a metal gate, especially in mixed signal circuits. Control of metal gate work-functions will also be a critical factor in fully depleted-SOI (FD–SOI)

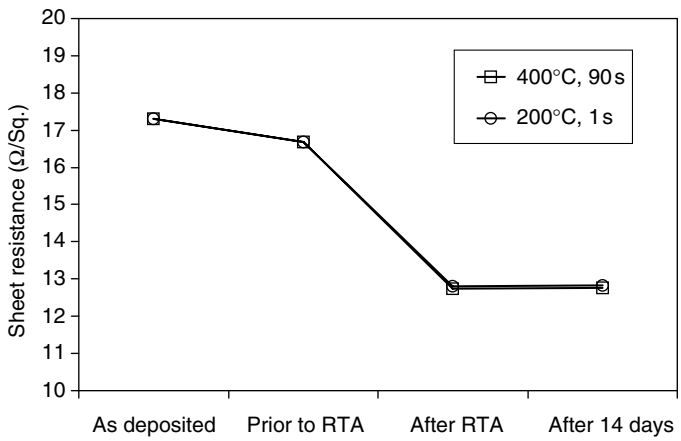


FIGURE 11.71 Dependence of sheet resistance of electroplated copper films on annealing temperature and time in RTP. (From Clarke, D., Bhaskaran, V., Sanchez, J., Broadbent, E., and Thakur, R., In *7th International Conference on Advanced Thermal Processing of Semiconductors—RTP 1999*, edited by Kitayama, H., Lojek, B., Miner, G., and Tillmann, A., 113, RTP 99, 1999.)

and multi-gate device technologies since in these devices, doping does not provide a convenient approach for threshold voltage control [483].

The importance of metal gates for future semiconductor devices has stimulated a resurgence of work on the materials science of the gate electrode and on fabrication approaches. It has even been suggested that metal gate technology may bring significant benefits, even without a change to a high-κ dielectric. However, a change from polysilicon also brings severe challenges in materials science and for the process integration approach. The normal CMOS process flow, where *s*/*d* doping and activation is performed after gate formation, poses a serious difficulty with respect to thermal stability of the metal gate [484].

A basic problem with a switch to metal gates arises because, for bulk CMOS it is necessary to use gate electrodes with different work-functions for the NMOS and PMOS devices. Hence, the quest for suitable metals becomes a challenge of work-function engineering. The need to deposit and pattern two separate metals also adds process complexity, and there has been considerable emphasis on finding approaches that require fewer process steps. Several approaches rely on depositing one metal and then adjusting its work-function on one set of devices by a modification process, such as an ion implantation step or an alloying process [485–487].

One promising approach for forming metal gates is through the "full-silicidation" of a silicon gate ("FUSI" process) [488–493]. In this process, a metal layer is deposited over a polysilicon gate and a thermal treatment completely converts it to a silicide layer that acts as the gate electrode. The approach has been evaluated for $CoSi_2$, PtSi, and NiSi electrodes, although the latter has received more emphasis. By performing the process on a doped polysilicon layer, it is possible to create a pile-up of dopants near the gate dielectric interface. The nature of the doping can be used to adjust the gate work-function. This kind of capability may prove important for future metal-gate technologies. The silicidation process is usually performed with RTP, and it shares many aspects with the NiSi processes discussed above [397,491].

This challenge of work-function engineering presents an opportunity where the excellent thermal and ambient process control of RTP can be very important. Process control will be crucial because the work-function of the gate electrode has a direct impact on the threshold voltage of the MOS device. Processing of metals is also likely to require the same very tight ambient control capabilities that have been typically associated with silicide processing requirements.

Changes in device architecture will also affect thermal processing needs. Applications where RTP processing enables optimization of stress and topography can be expected to expand in the era of 3D CMOS devices, where engineering the corners and smoothing the surfaces of silicon fins and other multi-gate structures will be critical [494–497]. Finally, the possibility of non-silicon CMOS has opened up new challenges in applying RTP to processing of other materials. For example, RTP methods have recently been applied to formation of junctions and contacts in germanium [498,499].

11.3.5 Applications of RTP beyond CMOS

As well as serving in the three main sectors of dielectrics, implant annealing and contact formation, RTP has found a variety of other uses in silicon device fabrication, as well as, a host of other applications outside the field of mainstream CMOS device technology. Table 11.1b suggests some of the fields where RTP plays an important role, and doubtless new applications will evolve wherever there is a need for precisely controlled thermal processing of high-value materials.

11.4 Conclusion

Rapid thermal processing technology has reached a high degree of maturity and is in widespread use in semiconductor device manufacturing. It provides the most flexible and cost-effective thermal processing technique for fabrication of advanced semiconductor devices on large-diameter wafers. The key enabling features are the rapid response of the heating energy source, the relatively cool chamber walls, the use of closed-loop temperature control and the very clean gas ambient. The original problems of wafer

temperature measurement and process uniformity control have been resolved and advanced RTP systems are capable of process control that is better than that in batch furnace processing. The advent of 300 mm wafers has accelerated the deployment of RTP as the preferred method for thermal processing for a combination of reasons including the need for restricted thermal budgets, reduced thermal stresses, and fast manufacturing cycle-times. Any further increases in wafer diameter will favor the continued adoption of RTP and RTCVD technology. As thermal processing continues its migration to RTP, the focus of equipment development is likely to shift towards increasing productivity through throughput improvements and simplification of operating and set-up procedures.

As semiconductor devices continue to scale, the need for reduced thermal budget will continue to drive applications towards RTP approaches. The great reduction in cycle-time, as compared to traditional batch furnace approaches, is also a major help for device manufacturers who need to handle a rapidly changing product mix. New applications are driving the introduction of many new process gas ambients and further improvement of ambient purity control. We can also expect to see RTP techniques extend further into film deposition, where RTCVD is already an important approach. Furthermore, rapid, well-controlled heating, combined with flexibility in ambient selection and control, can benefit a wide variety of applications in processing of high-technology electronic, optical and magnetic materials, and devices.

Acknowledgments

I would like to thank my colleagues at Mattson Technology for their help with the preparation of this chapter, especially Z. Nényei, J. Niess, W. Lerch, S. Paul, R. Sharangpani, S. McCoy and J. Gelpey, who contributed many of the process results described here.

References

1. Roozeboom, F., ed. *Advances in Rapid Thermal and Integrated Processing*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1996.
2. Fair, R. B., ed. *Rapid Thermal Processing*. San Diego, CA: Academic Press, 1993.
3. Lindholm, D. In *Rapid Thermal Processing '97*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 25–7. Round Rock, TX: RTP '97, 1997.
4. Singh, R. *J. Appl. Phys.* 63 (1988): R59.
5. Tillmann, A., S. Buschbaum, S. Frigge, U. Kreiser, D. Löffelmacher, T. Theilig, and P. Schmid. *Mater. Sci. Semicond. Process.* 1 (1998): 181.
6. Moslehi, M. M., L. Velo, A. Paranjpe, J. Kuehne, S. Huang, C. Schaper, T. Breedijk., et al. In *Rapid Thermal Processing '93*, edited by R. B. Fair, and B. Lojek, 43–59. Scottsdale, AZ: RTP '93, 1993.
7. Peuse, B., G. Miner, M. Yam, and C. Elia. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 71.
8. Gat, A., Z. Koren, P. J. Timans, and R. P. S. Thakur. In *Advances in Rapid Thermal Processing*, edited by F. Roozeboom, J. Gelpey, M. C. Öztürk, J. Nakos, and M. D. Allendorf, Seattle, WA: Proceedings of the Electrochemical Society Meeting, 1999.
9. Camm, D. M., M. Lefrançois, B. Hickson, D. Parfeniuk, and B. Lojek. In *Rapid Thermal Processing '95*, edited by R. B. Fair, and B. Lojek, 241–4. Round Rock, TX: RTP '95, 1995.
10. Shimizu, A., F. Mieno, A. Tsukune, H. Nomura, H. Ohta, H. Tokunoh, M. Kuramae, N. Setoguti, K. Watanabe, and Y. Furumura. In *Rapid Thermal Processing '93*, edited by R. B. Fair, and B. Lojek, 324–8. Scottsdale, AZ: RTP '93, 1993.
11. Yoo, W. S., and A. J. Atanos. In *Rapid Thermal Processing '98*, edited by T. Hori, B. Lojek, Y. Tanabe, and R. P. S. Thakur, 21–8. Round Rock, TX: RTP '98, 1998.
12. Lee, C., and A. B. Wittkower. In *Rapid Thermal Processing '93*, edited by R. B. Fair, and B. Lojek, 451–4. Scottsdale, AZ: RTP '93, 1993.
13. Timans, P. J. *Mater. Sci. Semicond. Process.* 1 (1998): 169.
14. Timans, P. J. In *Advances in Rapid Thermal and Integrated*, edited by F. Roozeboom, 35–102. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1996.

15. Siegel, R., and J. R. Howell. *Thermal Radiation Heat Transfer*. 3rd ed., 22–37. Washington, DC: Hemisphere Publishing Corporation, 1992.
16. Siegel, R., and J. R. Howell. *Thermal Radiation Heat Transfer*. 3rd ed., 47–86. Washington, DC: Hemisphere Publishing Corporation, 1992.
17. Hecht, E., and A. Zajac. *Optics*. 29–59. Reading, MA: Addison-Wesley, 1974.
18. Macleod, H. A. *Thin Film Optical Filters*. 67–8. New York: McGraw-Hill, 1986.
19. Hebb, J. P., K. F. Jensen, and J. Thomas. *IEEE Trans. Semicond. Manuf.* 11 (1998): 607.
20. Edwards, D. F. In *Handbook of Optical Constants of Solids*, edited by E. D. Palik, 547–69. San Diego, CA: Academic Press, 1985.
21. Aspnes, D. E. *Properties of Silicon*. 59–79. London: INSPEC, The Institution of Electrical Engineers, 1988.
22. Jellison, G. E. Jr., and F. A. Modine. *J. Appl. Phys.* 76 (1994): 3758.
23. Rogne, H., P. J. Timans, and H. Ahmed. *Appl. Phys. Lett.* 69 (1996): 2190.
24. Timans, P. J. *J. Appl. Phys.* 74 (1993): 6353.
25. Macfarlane, G. G., T. P. McLean, J. E. Quarrington, and V. Roberts. *Phys. Rev.* 111 (1958): 1245.
26. Jellison, G. E. Jr., and D. H. Lowndes. *Appl. Phys. Lett.* 41 (1982): 594.
27. Sturm, J. C., and M. Reaves. *IEEE Trans. Electron Devices* 39 (1992): 81.
28. Vandenabeele, P., and K. Maex. *J. Appl. Phys.* 72 (1992): 5867.
29. Boyd, I. W., T. D. Binnie, J. I. B. Wilson, and M. J. Colles. *J. Appl. Phys.* 55 (1984): 3061.
30. Sato, T. *Jpn. J. Appl. Phys.* 6 (1967): 339.
31. Stierwalt, D. L., and R. F. Potter. *J. Phys. Chem. Solids* 23 (1962): 99.
32. Bendow, B., H. G. Lipson, and P. Yukon. *Appl. Opt.* 16 (1977): 2909.
33. Li, H. H. *J. Phys. Chem. Ref. Data* 9 (1980): 561.
34. Magunov, A. N. *Opt. Spectrosc.* 73 (1992): 205.
35. Schumann, P. A. Jr., W. A. Keenan, A. H. Tong, H. H. Gegenwarth, and P. Schneider. *J. Electrochem. Soc.* 118 (1971): 145.
36. Soref, R. A., and B. R. Bennett. *IEEE J. Quantum Electron.* QE-23 (1987): 123.
37. Philipp, H. R. *Properties of Silicon*. 1015–27. London: INSPEC, The Institution of Electrical Engineers, 1988.
38. Philipp, H. R. *Properties of Silicon*. 1031–36. London: INSPEC, The Institution of Electrical Engineers, 1988.
39. Smith, D. Y., E. Shiles, and M. Inokuti. In *Handbook of Optical Constants of Solids*, edited by E. D. Palik, 369–406. San Diego, CA: Academic Press, 1985.
40. Ordal, M. A., L. L. Long, R. J. Bell, S. E. Bell, R. R. Bell, R. W. Alexander Jr., and C. A. Ward. *Appl. Opt.* 22 (1983): 1099.
41. Ward, L. In *Handbook of Optical Constants of Solids II*, edited by E. D. Palik, 435–48. San Diego, CA: Academic Press, 1991.
42. Gushchin, V. S., K. M. Shvarev, B. A. Baum, and P. V. Geld. *Sov. Phys. Dokl.* 23 (1978): 344.
43. Lynch, D. W., and W. R. Hunter. In *Handbook of Optical Constants of Solids*, edited by E. D. Palik, 280–6. San Diego, CA: Academic Press, 1985.
44. Mash, I. D., and G. P. Motulevich. *Sov. Phys. JETP* 36 (1973): 516.
45. Lynch, D. W., and W. R. Hunter. In *Handbook of Optical Constants of Solids*, edited by E. D. Palik, 357–67. San Diego, CA: Academic Press, 1985.
46. Wu, Z.-C., E. T. Arakawa, J. R. Jimenez, and L. J. Schowalter. *Phys. Rev. B* 47 (1993): 4356.
47. Wölfel, M., M. Schulz, J. Ionally, and P. J. Grunthaner. *Appl. Phys. A* 50 (1990): 177.
48. Duboz, J. Y., P. A. Badoz, J. Henz, and H. von Känel. *J. Appl. Phys.* 68 (1990): 2346.
49. Tanaka, M., S. Kurita, M. Fujisawa, and F. Lévy. *Phys. Rev. B* 43 (1991): 9133.
50. Borghesi, A., A. Piaggi, G. Guizzetti, F. Lévy, M. Tanaka, and H. Fukutani. *Phys. Rev. B* 40 (1989): 1611.
51. Lee, K., and J. T. Lue. *Phys. Lett. A* 125 (1987): 271.
52. Henrion, W., and H. Lange. *Phys. Status Solidi (b)* 151 (1989): 375.
53. Valkonen, E., C.-G. Ribbing, and E. Sundgren. *Proc. SPIE* 652 (1986): 235.
54. Timans, P. J. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 3.

55. Liebert, C. H. "Thermophysics of Spacecraft and Planetary Bodies." *Progress in Astronautics and Aeronautics*. Vol. 20, 17–40, 1967.
56. Timans, P. J. In *Rapid Thermal Processing '93*, edited by R. B. Fair, and B. Lojek, 282–6. Scottsdale, AZ: RTP '93, 1993.
57. Timans, P. J. In *Rapid Thermal Processing '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 145–56. Round Rock, TX: RTP '96, 1996.
58. Hill, C., S. Jones, and D. Boys. In *Reduced Thermal Processing for ULSI*, edited by R. A. Levy, 143–80. New York: Plenum Press, 1989.
59. Rogne, H., and H. Ahmed. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 27.
60. Timans, P. J. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 186–93. Round Rock, TX: RTP '94, 1994.
61. Pettibone, D. W., J. R. Suarez, and A. Gat. *Mater. Res. Soc. Symp. Proc.* 52 (1986): 209.
62. Nulman, J., S. Antonio, and W. Blonigan. *Appl. Phys. Lett.* 56 (1990): 2513.
63. Nulman, J., B. Cohen, W. Blonigan, S. Antonio, R. Meinecke, and A. Gat. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 461.
64. Nakos, J. S. In *Rapid Thermal Processing '93*, edited by R. B. Fair, and B. Lojek, 421–8. Scottsdale, AZ: RTP '93, 1993.
65. Delfino, M., and T. Hodul. *IEEE Trans. Electron Devices* 39 (1992): 89.
66. Ravindra, N. M., F. M. Tong, W. F. Kosonocky, J. R. Markham, S. Liu, and K. Kinsella. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 431.
67. Schietinger, C. In *Rapid Thermal Processing '95*, edited by R. B. Fair, and B. Lojek, 225–33. Round Rock, TX: RTP '95, 1995.
68. Chr, P. L. A., M. van der Meer, L. J. Giling, and G. Kroon. *J. Appl. Phys.* 47 (1976): 652.
69. Nulman, J. *Proc. SPIE* 1189 (1989): 72.
70. Sturm, J. C., and A. Reddy. *Mater. Res. Soc. Symp. Proc.* 387 (1995): 137.
71. Vandenabeele, P., R. J. Schreutelkamp, K. Maex, C. Vermeiren, and W. Coppye. *Mater. Res. Soc. Symp. Proc.* 260 (1992): 653.
72. Vandenabeele, P., and K. Maex. *Proc. SPIE* 1189 (1989): 89.
73. Öztürk, M. C., M. K. Sanganeria, and Y. Sorrell. *Appl. Phys. Lett.* 61 (1992): 2697.
74. Wong, P. Y., C. K. Hess, and N. Miaoulis. *Mater. Res. Soc. Symp. Proc.* 303 (1993): 217.
75. Wong, P. Y., and N. Miaoulis. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 395.
76. Vandenabeele, P., and K. Maex. *Proc. SPIE* 1393 (1990): 316.
77. Takeuti, D. F., P. J. Timans, and H. Ahmed. *Appl. Phys. Lett.* 67 (1995): 2206.
78. Xu, H., and J. C. Sturm. *Mater. Res. Soc. Symp. Proc.* 387 (1995): 29.
79. Lord, H. A. *IEEE Trans. Semicond. Manuf.* 1 (1988): 105.
80. Kakoschke, R., and E. Bußmann. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 473.
81. Cho, Y. M., A. Paulraj, S. Norman, G. Xu, and T. Kailath. *Proc. SPIE* 1804 (1992): 34.
82. Zöllner, J.-P., I. Patzschke, V. Pietzuch, J. Pezoldt, and G. Eichhorn. *Mater. Res. Soc. Symp. Proc.* 303 (1993): 177.
83. Dilhac, J.-M., and N. Nolhier. In *Rapid Thermal Processing '93*, edited by R. B. Fair, and B. Lojek, 12–21. Scottsdale, AZ: RTP '93, 1993.
84. Knutson, K. L., T. P. Merchant, J. V. Cole, J. P. Hebb, T. G. Mihopuolos, and K. F. Jensen. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 146–52. Round Rock, TX: RTP '94, 1994.
85. Ting, A. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 102–9. Round Rock, TX: RTP '94, 1994.
86. Spence, P. A., W. S. Winters, R. J. Kee, and A. Kermani. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 139–45. Round Rock, TX: RTP '94, 1994.
87. Spence, P., C. Schaper, and A. Kermani. *Mater. Res. Symp. Proc.* 387 (1995): 75.
88. Ebert, J. L., A. Emami-Naeini, and R. Kosut. In *Rapid Thermal Processing '95*, edited by R. B. Fair, and B. Lojek, 343–55. Round Rock, TX: RTP '95, 1995.
89. Kersch, A., Th.Schafbauer, H.-J. Timme, and A. Ajmera. In *Rapid Thermal Processing '95*, edited by R. B. Fair, and B. Lojek, 367–75. Round Rock, TX: RTP '95, 1995.

90. Kersch, A., and T. Schafbauer. In *Rapid Thermal Processing '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 347–55. Round Rock, TX: RTP '96, 1996.
91. Schafbauer, T., and A. Kersch. In *Rapid Thermal Processing '97*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 298–304. Round Rock, TX: RTP '97, 1997.
92. Fordham, M., M. Pan, J. Hu, and Y. Sorrell. *Mater. Res. Symp. Proc.* 470 (1997): 175.
93. Tillmann, A. In *Rapid Thermal Processing '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 356–71. Round Rock, TX: RTP '96, 1996.
94. Erofeev, A. F., A. V. Kolpakov, T. M. Makhviladze, A. V. Martjushenko, A. V. Panjukhin, O. S. Volchek, and M. Orlowski. In *Rapid Thermal Processing '95*, edited by R. B. Fair, and B. Lojek, 181–97. Round Rock, TX: RTP '95, 1995.
95. Jensen, K. F., T. P. Merchant, J. V. Cole, J. P. Hebb, K. L. Knutson, and T. G. Mihopoulos. In *Advances in Rapid Thermal and Integrated Processing*, edited by F. Roozeboom, 265–304. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1996.
96. Jensen, K. F., H. Simka, T. G. Mihopoulos, P. Futerko, and M. Hierlemann. In *Advances in Rapid Thermal and Integrated Processing*, edited by F. Roozeboom, 305–31. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1996.
97. Balakrishnan, K. S., T. L. Cooper, and T. F. Edgar. In *Rapid Thermal Processing '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 279–86. Round Rock, TX: RTP '96, 1996.
98. Edgar, T. F., and T. Breedijk. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 266–77. Round Rock, TX: RTP '94, 1994.
99. Balakrishnan, K. S., S. Shooshtarian, N. Acharya, P. J. Timans, and R. P. S. Thakur. In *Advances in Rapid Thermal Processing*, edited by F. Roozeboom, J. Gelpey, M. C. Öztürk, J. Nakos, and M. D. Allendorf, Proceedings of the Electrochemical Society Meeting, Seattle, Washington, 1999.
100. Nutter, G. D. In *Theory and Practice of Radiation Thermometry*, edited by D. P. DeWitt, and G. D. Nutter, 231–337. New York: Wiley, 1988.
101. Timans, P. J. *Solid State Technol.* 40 (1997): 63.
102. Walk, H., and T. Theiler. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 194–6. Round Rock, TX: RTP '94, 1994.
103. Vandenabeele, P., and W. Renken. *Mater. Res. Soc. Symp. Proc.* 470 (1997): 17.
104. Vandenabeele, P., and W. Renken. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 115.
105. Renken, W. In *Rapid Thermal Processing '93*, edited by R. B. Fair, and B. Lojek, 262–6. Scottsdale, AZ: RTP '93, 1993.
106. Kreider, K. G., D. P. DeWitt, B. K. Tsai, F. J. Lovas, and D. W. Allen. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 87.
107. Yam, M., A. Rubinchik, and B. Peuse. In *Rapid Thermal Processing '97*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 102–4. Round Rock, TX: RTP '97, 1997.
108. Dilhac, J.-M., C. Ganibal, N. Nolhier, and B. Rousset. *Rev. Sci. Instrum.* 63 (1992): 188.
109. Roozeboom, F., and N. Parekh. *J. Vac. Sci. Technol. B* 8 (1990): 1249.
110. See, A., H. Wong, Y.-S. Lin, and L.-H. Chua. In *Rapid Thermal Processing '97*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 118–23. Round Rock, TX: RTP '97, 1997.
111. Stein, A. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 178–81. Round Rock, TX: RTP '94, 1994.
112. Lefrançois, M. E., D. M. Camm, and J. Hickson. *Mater. Res. Symp. Proc.* 429 (1996): 321.
113. Fiory, A. T., C. Schietinger, B. Adams, and G. Tinsley. *Mater. Res. Soc. Symp. Proc.* 303 (1993): 139.
114. Fiory, A. T., and K. Nanda. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 3.
115. Oh, M., B. Nguyenphu, and T. Fiory. *Mater. Res. Soc. Symp. Proc.* 387 (1995): 131.
116. Nguyenphu, B., M. Oh, and T. Fiory. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 291.
117. Glazman, E., P. Alezra, Z. Atzmon, H. Gilboa, and A. Thon. In *Rapid Thermal Processing '98*, edited by T. Hori, B. Lojek, Y. Tanabe, and R. P. S. Thakur, 146–55. Round Rock, TX: RTP '98, 1998.
118. Kruwinus, H. In *Rapid Thermal Processing '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 85–7. Round Rock, TX: RTP '96, 1996.
119. Liu, E. Z., H. M. Li, H. L. Chew, Y. S. Lin, and L. Chan. In *Rapid Thermal Processing '95*, edited by R. B. Fair, and B. Lojek, 210–3. Round Rock, TX: RTP '95, 1995.

120. Lang, M. K., G. W. Donohoe, S. H. Zaidi, and R. J. Brueck. *Opt. Eng.* 33 (1994): 3465.
121. Peuse, B., and A. Rosekrans. *Mater. Res. Soc. Symp. Proc.* 303 (1993): 125.
122. Degertekin, F. L., J. Pei, B. T. Khuri-Yakub, and K. Saraswat. *Appl. Phys. Lett.* 64 (1994): 1338.
123. Klimek, D., B. Anthony, A. Abbate, and P. Kotidis. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 135.
124. Adel, M. E., Y. Ish-Shalom, S. Mangan, D. Cabib, and H. Gilboa. *Proc. SPIE* 1803 (1992): 290.
125. Sturm, J. C., P. V. Schwartz, and M. Garone. *Appl. Phys. Lett.* 56 (1990): 961.
126. Cullen, C. W., and C. Sturm. *IEEE Trans. Semicond. Manuf.* 8 (1995): 346.
127. Guidotti, D. *J. Vac. Sci. Technol. B* 16 (1998): 609.
128. Massoud, H. Z. In *Rapid Thermal Processing '93*, edited by R. B. Fair, and B. Lojek, 267–73. Scottsdale, AZ: RTP '93, 1993.
129. Donnelly, V. M., and J. A. McCaulley. *J. Vac. Sci. Technol. A* 8 (1990): 84.
130. Lo, H. W., and A. Compaan. *J. Appl. Phys.* 51 (1980): 1565.
131. Jellison, G. E. Jr., D. H. Lowndes, and R. F. Wood. *Phys. Rev. B* 28 (1983): 3272.
132. Vandenabeele, P., and W. Renken. *Mater. Res. Soc. Symp. Proc.* 470 (1997): 181.
133. Hayn, R., A. Tillmann, and W. Kegel. In *Rapid Thermal Processing '97*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 311–20. Round Rock, TX: RTP '97, 1997.
134. Timme, H.-J., T. Nguyen, and A. Ajmera. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 314–20. Round Rock, TX: RTP '94, 1994.
135. Li, J. G., R. J. Champetier, and P. J. Timans. In *Rapid Thermal Processing '97*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 305–10. Round Rock, TX: RTP '97, 1997.
136. Nenyei, Z., A. Gschwandtner, and S. Marcus. In *Rapid Thermal Processing '95*, edited by R. B. Fair, and B. Lojek, 58–68. Round Rock, TX: RTP '95, 1995.
137. Timans, P. J., R. N. Morishige, and Y. Wasserman. *Mater. Res. Soc. Symp. Proc.* 470 (1997): 57.
138. Christ, R. S. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 60–4. Round Rock, TX: RTP '94, 1994.
139. Pfeifer, K., and B. Roche. In *Rapid Thermal Processing '95*, edited by R. B. Fair, and B. Lojek, 319–24. Round Rock, TX: RTP '95, 1995.
140. Lerch, W., W. Blersch, and S. Yanagawa. *IEEE Trans. Semicond. Manuf.* 11 (1998): 598.
141. Dilhac, J.-M., C. Ganibal, and A. Martinez. *Mater. Res. Soc. Symp. Proc.* 92 (1987): 259.
142. Nenyei, Z., A. Tillmann, and J. Gelpey. In *Rapid Thermal Processing '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 237–45. Round Rock, TX: RTP '96, 1996.
143. Koutny, W. In *Rapid Thermal Processing '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 272–8. Round Rock, TX: RTP '96, 1996.
144. Shi, J., R. J. Bradley, and L. A. Larson. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 321–4. Round Rock, TX: RTP '94, 1994.
145. Hodul, D., and S. Mehta. *Mater. Res. Soc. Symp. Proc.* 92 (1987): 183.
146. Widenhofer, R. D., S. D. Marcus, and S. K. Pozder. In *Rapid Thermal Processing '97*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 212–6. Round Rock, TX: RTP '97, 1997.
147. Dilhac, J.-M. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 333.
148. Lange, P., E. Hartmannsgruber, and F. Naumann. In *Rapid Thermal Processing '94*, edited by R. B. Fair, and B. Lojek, 219–26. Round Rock, TX: RTP '94, 1994.
149. Vandenabeele, P. PhD dissertation, 45–6. Catholic University of Leuven, 1994.
150. Vandenabeele, P. PhD dissertation, 27. Catholic University of Leuven, 1994.
151. Tsai, Y.-T., R. Subrahmanyan, A. R. Sitaram, and M. Orlowski. *Mater. Res. Soc. Symp. Proc.* 303 (1989): 231.
152. Leitz, G., J. Pezoldt, I. Patzschke, J.-P. Zöllner, and G. Eichhorn. *Mater. Res. Soc. Symp. Proc.* 303 (1989): 171.
153. Blake, J., J. C. Gelpey, J. F. Moquin, J. Schlueter, and R. Capodilupo. *Mater. Res. Soc. Symp. Proc.* 92 (1987): 265.
154. Moslehi, M. M. *IEEE Trans. Semicond. Manuf.* 4 (1989): 130.
155. Buller, J. F., M. M. Farahani, and S. Garg. *IEEE Trans. Semicond. Manuf.* 9 (1996): 108.
156. Knutson, K. L., and T. L. Cooper. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 31.

157. Acharya, N., V. Kirtikar, S. Shooshtarian, H. Doan, P. J. Timans, K. S. Balakrishnan, and L. Knutson. *IEEE Trans. Semicond. Manuf.* 14 (2001): 218.
158. Tillmann, A., and T. Knarr. In *Rapid Thermal Processing '95*, edited by R. B. Fair, and B. Lojek, 214–20. Round Rock, TX: RTP '95, 1995.
159. Vosen, S. R., P. Timans, J. Li, and N. Acharya. In *7th International Conference on Advanced Thermal Processing of Semiconductors—RTP '99*, edited by H. Kitayama, B. Lojek, G. Miner, and A. Tillmann, 281. Colorado Springs, CO: RTP '99, 1999.
160. Peuse, B., M. Pfarr, P. Timans, and Y. Hu. In *12th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2004*, edited by J. Gelpey, B. Lojek, Z. Nenyei, and R. Singh, 61. Piscataway, NJ: IEEE, 2004.
161. De Roover, D., A. Emami-Naeini, J. L. Ebert, S. Ghosal, and G. van der Linden. In *Rapid Thermal Processing '98*, edited by T. Hori, B. Lojek, Y. Tanabe, and R. P. S. Thakur, 177–86. Round Rock, TX: RTP '98, 1998.
162. Schaper, C. D., T. Kailath, and Y. J. Lee. *IEEE Trans. Semicond. Manuf.* 12 (1999): 193.
163. Sorrell, F. Y., J. A. Harris, M. C. Ozturk, and J. J. Wortman. *Proc. SPIE* 1189 (1989): 30.
164. Liao, J. C., and T. I. Kamins. *J. Appl. Phys.* 67 (1990): 3848.
165. Vandenabeele, P., K. Maex, and R. De Keersmaecker. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 149.
166. Hebb, J. P., and K. F. Jensen. *J. Electrochem. Soc.* 143 (1996): 1142.
167. Hebb, J. P., and K. F. Jensen. *Mater. Res. Symp. Proc.* 429 (1996): 43.
168. Kersch, A., T. Schafbauer, and L. Deutschmann. *Mater. Res. Symp. Proc.* 429 (1996): 71.
169. Bremensdorfer, R., S. Marcus, and Z. Nenyei. *Mater. Res. Symp. Proc.* 429 (1996): 327.
170. Kersch, A. *Mater. Res. Symp. Proc.* 470 (1997): 159.
171. Erofeev, A. F., T. M. Makhviladze, A. V. Panjukhin, O. S. Volchek, and O. Adetutu. In *Rapid Thermal Processing '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 342–6. Round Rock, TX: RTP '96, 1996.
172. Lefrancois, M. E., D. M. Camm, and B. J. Hickson. *Mater. Res. Symp. Proc.* 429 (1996): 321.
173. Kuehne, J., S. Hattangady, and M. Pas. In *4th International Conference on Advanced Thermal Processing of Semiconductors—RTP '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 417. Round Rock, TX: RTP '96, 1996.
174. Timans, P. J., Z. Nényei, and R. Berger. *Solid State Technol.* 45 (2002): 67.
175. Aderhold, W., S. Poarch, and A. Hunter. In *10th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2002*, edited by J. Gelpey, B. Lojek, Z. Nényei, and R. Singh, 69. Piscataway, NJ: IEEE, 2002.
176. Niess, J., R. Berger, P. J. Timans, and Z. Nényei. In *10th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2002*, edited by J. Gelpey, B. Lojek, Z. Nenyei, and R. Singh, 49. Piscataway, NJ: IEEE, 2002.
177. Berger, R., S. Miethaner, H. Gruber, J. Niess, W. Dietl, and Z. Nényei. In *9th International Conference on Advanced Thermal Processing of Semiconductors—RTP 2001*, edited by D. P. DeWitt, J. Gelpey, B. Lojek, and Z. Nenyei, 72. Piscataway, NJ: IEEE, 2001.
178. Niess, J., Z. Nényei, W. Lerch, and S. Paul. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices*, edited by F. Roozeboom, E. P. Gusev, L. J. Chen, M. C. Öztürk, D.-L. Kwong, and P. J. Timans, 11. Pennington, NJ: The Electrochemical Society, 2003.
179. MacKnight, R. B., P. J. Timans, S.-P. Tay, and Z. Nenyei. In *12th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2004*, edited by J. Gelpey, B. Lojek, Z. Nenyei, and R. Singh, 3. Piscataway, NJ: IEEE, 2004.
180. Mattson, B., P. Timans, S.-P. Tay, D. J. Devine, and J. Kim. In *9th International Conference on Advanced Thermal Processing of Semiconductors—RTP 2001*, edited by D. P. DeWitt, J. Gelpey, B. Lojek, and Z. Nenyei, 13. Piscataway, NJ: IEEE, 2001.
181. Mercer, D. E., A. Jain, and S. Watts Butler. In *Rapid Thermal and Other Short-Time Processing Technologies II*, edited by D.-L. Kwong, K. G. Reid, M. C. Öztürk, P. J. Timans, and F. Roozeboom, 247. Pennington, NJ: The Electrochemical Society, 2001.
182. Nényei, Z., H. Sommer, J. Gelpey, and A. Bauer. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 401.

183. Kondoh, E., G. Vereecke, M. M. Heyns, K. Maex, T. Gutt, and Z. Nényei. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 51.
184. Lerch, W., M. Glück, N. A. Stolwijk, H. Walk, M. Schäfer, S. D. Marcus, D. F. Downey, and J. W. Chow. *J. Electrochem. Soc.* 146 (1999): 2670.
185. Chen, C., T. Lin, J. Jung, N. Yabuoshi, Y. Sasaki, K. Komori, and H. H. Shih, et al. In *2001 International Electron Devices Meeting Technical Digest*, 28.3.1. Piscataway, NJ: IEEE, 2001.
186. Ma, Y. In *Rapid Thermal and Other Short-Time Processing Technologies II*, edited by D.-L. Kwong, K. G. Reid, M. C. Öztürk, P. J. Timans, and F. Roozeboom, 3. Pennington, NJ: The Electrochemical Society, 2001.
187. Nagabushnam, R. V., R. K. Singh, and S. Sharan. *Mater. Sci. Semicond. Process.* 1 (1998): 207.
188. Fair, R. B. In *Rapid Thermal and Other Short-Time Processing Technologies*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, K. Reid, and D.-L. Kwong, 21. Pennington, NJ: The Electrochemical Society, 2000.
189. Chen, C.-C., V. S. Chang, Y. Jin, C.-H. Chen, T.-L. Lee, S.-C. Chen, and M.-S. Liang. In *2004 Symposium on VLSI Technology Technical Digest*, 176. Piscataway, NJ: IEEE, 2004.
190. Fukano, A., and H. Oyanagi. *Mater. Res. Soc. Symp.* 811 (2004): E1.3.1.
191. Huang, C.-H., and J.-G. Hwu. *Solid-State Electron.* 44 (2000): 1405.
192. Borisenko, V. E., and P. J. Hesketh. *Rapid Thermal Processing of Semiconductors*. New York: Plenum Press, 1997.
193. Nulman, J. In *Reduced Thermal Processing for ULSI*, edited by R. A. Levy, New York: Plenum Press, 1989.
194. Moslehi, M. M. *Appl. Phys. A* 46 (1988): 255.
195. Green, M. L. "Rapid Thermal O_2-Oxidation and N_2O Oxidation." In *Advances in Rapid Thermal and Integrated Processing*, edited by F. Roozeboom, 193. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1996 Chapter 7.
196. Weimer, R. A., D. M. Eppich, K. L. Beaman, D. C. Powell, and F. Gonzalez. *IEEE Trans. Semicond. Manuf.* 16 (2003): 138.
197. Sharangpani, R., J. H. Das, and S.-P. Tay. In *Rapid Thermal and Other Short-Time Processing Technologies*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, K. Reid, and D.-L. Kwong, 203. Pennington, NJ: The Electrochemical Society, 2000.
198. Chen, K.-C., H.-H. Shih, Y.-L. Hwang, C.-C. Hsueh, H. Chung, S. Pan, and C. Y. Lu. *IEEE Trans. Semicond. Manuf.* 16 (2003): 128.
199. Joshi, A., and D. L. Kwong. *IEEE Trans. Electron Devices* 39 (1992): 2099.
200. Joshi, A., and D. L. Kwong. *IEEE Electron Device Lett.* 12 (1991): 28.
201. Fonseca, L., and F. Campabadal. *IEEE Electron Device Lett.* 15 (1994): 449.
202. Arakawa, T., H. Fukuda, Y. Okabe, T. Iwabuchi, and S. Ohno. *J. Electrochem. Soc.* 137 (1990): 1650.
203. Yoneda, K., Y. Todokoro, and M. Inoue. *J. Mater. Res.* 6 (1991): 2362.
204. Inoue, M., and K. Yoneda. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 283.
205. Ludsteck, A., J. Schulze, I. Eisele, W. Dietl, and Z. Nenyei. *J. Appl. Phys.* 95 (2004): 2827.
206. Fiory, A. T. *Mater. Res. Soc. Symp. Proc.* 567 (1999): 13.
207. Moslehi, M. M., S. C. Shatas, and K. C. Saraswat. *Appl. Phys. Lett.* 47 (1985): 1353.
208. Deaton, R., and H. Z. Massoud. *Mater. Res. Soc. Symp. Proc.* 224 (1991): 373.
209. Yahia-Messaoud, A., G. Sarrabayrouse, A. Claverie, A. Martinez, E. Scheid, E. Campo, and M. Faye. *Mater. Res. Soc. Symp. Proc.* 224 (1991): 391.
210. Paz de Araujo, C. A., R. W. Gallegos, and Y. P. Huang. *J. Electrochem. Soc.* 136 (1989): 2673.
211. Chiou, Y. L., C. H. Sow, G. Li, and A. Ports. *Appl. Phys. Lett.* 57 (1990): 881.
212. Lassig, S. E., and J. L. Crowley. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 307.
213. Deaton, R. In *1st International Rapid Thermal Processing Conference RTP '93*, edited by R. B. Fair, and B. Lojek, 117. Scottsdale, AZ: RTP '93, 1993.
214. Smith, F. W., and G. Ghidini. *J. Electrochem. Soc.* 129 (1982): 1300.
215. Starodub, D., E. P. Gusev, E. Garfunkel, and T. Gustafsson. *Surf. Rev. Lett.* 6 (1999): 45.
216. Liehr, M., J. E. Lewis, and G. W. Rubloff. *J. Vac. Sci. Technol.* A 5 (1987): 1559.

217. Lerch, W., G. Roters, P. Munzinger, R. Mader, and R. Ostermeir. *Mater. Sci. Eng.* B54 (1998): 153.
218. Sharangpani, R., J. Das, S. P. Tay, R. P. S. Thakur, T. C. Yang, and C. Saraswat. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 143.
219. Lassig, S. E., T. J. Debolske, and L. Crowley. *Mater. Res. Soc. Symp. Proc.* 92 (1987): 103.
220. Grant, J. M., and Y. Hsieh. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 163.
221. Grant, J. M., and Z. Karim. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 257.
222. McIntosh, R., C. Galewski, and J. Grant. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 209.
223. Sun, S. C., L. S. Wang, F. L. Yeh, T. S. Lai, and H. Lin. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 181.
224. Ting, W., H. Hwang, J. Lee, and L. Kwong. *Appl. Phys. Lett.* 57 (1990): 2808.
225. Harrison, H. B., S. Dimitrijev, D. Sweatman, J. Parker, and S. Preston. *Mater. Res. Soc. Symp. Proc.* 303 (1993): 413.
226. Weidner, G., D. Krüger, M. Weidner, and K. Tittelbach-Helmrich. *Mater. Res. Soc. Symp. Proc.* 387 (1995): 265.
227. Yao, Z.-Q., H. B. Harrison, S. Dimitrijev, and T. Yeow. *IEEE Electron Device Lett.* 15 (1994): 516.
228. Sharangpani, R., S.-P. Tay, R. Thakur, S. Everist, J. Nelson, and P. M. Smith. In *Advances in Rapid Thermal Processing*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, and J. Nakos, 89. Pennington, NJ: The Electrochemical Society, 1999.
229. Harrison, H. B., Z.-Q. Yao, S. Dimitrijev, D. Sweatman, and T. Yeow. *Mater. Res. Soc. Symp. Proc.* 387 (1995): 233.
230. Kuehne, J., S. Hattangady, J. Piccirillo, G. C. Xing, G. E. Miner, and D. Lopes. *Mater. Res. Symp. Proc.* 470 (1997): 381.
231. Sharangpani, R., and P. Tay. *J. Electrochem. Soc.* 148 (2001): F5.
232. Gusev, E. P., H.-C. Lu, E. L. Garfunkel, T. Gustafsson, and L. Green. *IBM J. Res. Dev.* 43 (1999): 265.
233. Moslehi, M. M., and C. Saraswat. *IEEE Trans. Electron Devices* ED-32 (1985): 106.
234. Buchheit, K. M., H. Takeuchi, and T.-J. King. *Mater. Res. Soc. Symp. Proc.* 786 (2004): E2.2.1.
235. Green, M. L., T. Sorsch, L. C. Feldman, W. N. Lennard, E. P. Gusev, E. Garfunkel, H. C. Lu, and T. Gustafsson. *Appl. Phys. Lett.* 71 (1997): 2978.
236. Lu, Z. H., A. Khoueir, W. T. Ng, and S. P. Tay. In *Rapid Thermal and Other Short-Time Processing Technologies*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, K. Reid, and D.-L. Kwong, 223. Pennington, NJ: The Electrochemical Society, 2000.
237. Chang, K.-M., W.-C. Yang, C.-F. Chen, and B.-F. Hung. *J. Electrochem. Soc.* 151 (2004): F118.
238. Huff, H. R., G. A. Brown, L. A. Larson, and R. W. Murto. In *Rapid Thermal and Other Short-Time Processing Technologies II*, edited by D.-L. Kwong, K. G. Reid, M. C. Öztürk, P. J. Timans, and F. Roozeboom, 263. Pennington, NJ: The Electrochemical Society, 2001.
239. De Gendt, S., D. Brunco, M. Caymax, T. Conrad, L. Date, A. Delabie, W. Deweerd., et al. In *Advanced Gate Stack, Source/Drain and Channel Engineering for Si-Based CMOS: New Materials, Processes, and Equipment*, edited by E. P. Gusev, L. J. Chen, H. Iwai, D.-L. Kwong, M. C. Öztürk, F. Roozeboom, and P. J. Timans, 109. Pennington, NJ: The Electrochemical Society, 2005.
240. Eftekhari, G. *J. Electrochem. Soc.* 140 (1993): 787.
241. Yang, T. C., N. Bhat, and K. C. Saraswat. In *Proceedings of 4th Symposium on Silicon Nitride and Silicon Oxide Thin Insulating Films*, 191st Meeting of the Electrochemical Society, May 1997.
242. Jia, Y. B., J. Y. Choi, J. Schuur, J. H. Das, R. Sharangpani, and R. P. S. Thakur. In *Advances in Rapid Thermal Processing*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, and J. Nakos, 15. Pennington, NJ: The Electrochemical Society, 1999.
243. O'Sullivan, B. J., P. K. Hurley, C. Leveugle, and H. Das. *J. Appl. Phys.* 89 (2001): 3811.
244. O'Sullivan, B. J., P. K. Hurley, A. Mathewson, J. H. Das, and D. Daniel. *Microelectron. Reliability* 40 (2000): 645.
245. Itoh, S., G. Q. Lo, D. L. Kwong, V. K. Matthews, and C. Fazan. *IEEE Trans. Electron Devices* 40 (1993): 1176.
246. Cheung, A. W., G. Q. Lo, D.-L. Kwong, N. S. Alvi, and A. Kermani. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 483.

247. Chao, T. S., W. L. Yang, C.-M. Cheng, T. M. Pan, and T. F. Lei. In *2001 International Symposium on VLSI Technology, Systems, and Applications*, 142. Piscataway, NJ: IEEE, 2001.
248. Mattheus, A., A. Gschwandtner, R. Kakoschke, M. Kerber, and A. Talg. In *4th International Conference on Advanced Thermal Processing of Semiconductors—RTP '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 212. Round Rock, TX: RTP '96, 1996.
249. She, M., H. Takeuchi, and T. J. King. *IEEE Electron Device Lett.* 24 (2003): 309.
250. Yang, W.-C., C.-F. Chen, and K.-M. Chang. In *2003 IEEE Conference on Electron Devices and Solid-State Circuits*, 349. Piscataway, NJ: IEEE, 2004.
251. Mur, P., M. N. Semeria, M. Olivier, A. M. Papon, Ch. Leroux, G. Reimbold, P. Gentile., et al. *Appl. Surf. Sci.* 175–176 (2001): 726.
252. Bidaud, M., F. Guyader, F. Arnaud, J.-L. Autran, and J. Barla. *Non-Cryst. Solids* 280 (2001): 32.
253. Nayak, D. K., K. Kamjoo, J. S. Park, J. C. S. Woo, and K. L. Wang. *Appl. Phys. Lett.* 57 (1990): 369.
254. Nayak, D., K. Kamjoo, J. S. Park, J. C. S. Woo, and K. L. Wang. *IEEE Trans. Electron Devices* 39 (1992): 56.
255. Bera, L. K., W. K. Choi, C. S. Tan, S. K. Samanta, and C. K. Maiti. *IEEE Electron Device Lett.* 22 (2001): 387.
256. Chui, C. O., F. Ito, and K. C. Saraswat. *IEEE Electron Device Lett.* 25 (2004): 613.
257. Hames, G. A., S. E. Beck, A. G. Gilicinski, W. K. Henson, and J. J. Wortman. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 219.
258. Ahn, J., G. Q. Lo, W. Ting, D. L. Kwong, J. Kuehne, and C. W. Magee. *Appl. Phys. Lett.* 58 (1991): 425.
259. Glück, M., U. König, J. Hersener, Z. Nenyei, and A. Tillmann. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 215.
260. Tanabe, Y., Y. Nakatsuka, S. Sakai, T. Miyazaki, and T. Nagahama. *Proc. IEEE Int. Symp. Semicond. Manuf.* 1997 (1997): 49.
261. Reichenbach, D., R. Dubois, and J. Krasowski. In *8th International Conference on Advanced Thermal Processing of Semiconductors-RTP 2000*, edited by D. P. DeWitt, J. Kowalski, B. Lojek, and A. Tillmann, 132. Round Rock, TX: RTP Conference, 2000.
262. Beyer, A., R. Hayn, W. Kegel, and J.-U. Sachse. In *8th International Conference on Advanced Thermal Processing of Semiconductors-RTP 2000*, edited by D. P. DeWitt, J. Kowalski, B. Lojek, and A. Tillmann, 47. Round Rock, TX: RTP Conference, 2000.
263. Sharangpani, R., and S.-P. Tay. In *Rapid Thermal and Other Short-Time Processing Technologies II*, edited by D.-L. Kwong, K. G. Reid, M. C. Öztürk, P. J. Timans, and F. Roozeboom, 157. Pennington, NJ: The Electrochemical Society, 2001.
264. Reid, K. G., H. Tseng, R. Hedge, G. Miner, and G. Xing. In *Advances in Rapid Thermal Processing*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, and J. Nakos, 23. Pennington, NJ: The Electrochemical Society, 1999.
265. Sullivan, N., L. L. Raja, R. J. Kee, Y. Yokota, and M. Williams. In *9th International Conference on Advanced Thermal Processing of Semiconductors-RTP 2001*, edited by D. P. DeWitt, J. Gelpey, B. Lojek, and Z. Nenyei, 95. Piscataway, NJ: IEEE, 2001.
266. Das, J. H., C. Powell, V. Kirtikar, A. D. Daniel, R. Weimer, and S.-P. Tay. In *8th International Conference on Advanced Thermal Processing of Semiconductors-RTP 2000*, edited by D. P. DeWitt, J. Kowalski, B. Lojek, and A. Tillmann, 47. Round Rock, TX: RTP Conference, 2000.
267. Roters, G., R. Hayn, W. Kegel, O. Storbeck, S. Frigge, G. Feldmeyer, H. J. Meyer, and E. Schroer. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices*, edited by F. Roozeboom, E. P. Gusev, L. J. Chen, M. C. Öztürk, D.-L. Kwong, and P. J. Timans, 385. Pennington, NJ: The Electrochemical Society, 2003.
268. Liu, Y., and J. Hebb. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices*, edited by F. Roozeboom, E. P. Gusev, L. J. Chen, M. C. Öztürk, D.-L. Kwong, and P. J. Timans, 25. Pennington, NJ: The Electrochemical Society, 2003.
269. Lee, B. H., D. K. Sohn, J.-S. Park, C. H. Han, Y.-J. Huh, J. S. Byun, and J. J. Kim. *IEDM '98 Technical Digest*. 385. Piscataway, NJ: IEEE, 1998.

270. Thakur, R. P. S., S. J. DeBoer, H. N. Al-Shareef, D. Gealy, and R. Singh. In *Proceedings 2nd International Symposium on Low and High Dielectric Constant Materials: Materials Science, Processing and Reliability Issues*, Vol. 97-8, edited by H. S. Rathore, R. Singh, R. P. S. Thakur, and S. S. C. Sun, *Proceedings 2nd International Symposium on Low and High Dielectric Constant Materials: Materials Science, Processing and Reliability Issues*, 224. Pennington, NJ: The Electrochemical Society, Inc., 1997.
271. Green, M. L., T. W. Sorsch, G. Timp, E. L. Garfunkel, E. P. Gusev, T. Gustafsson, W. N. Lennard., et al. In *6th International Conference on Advanced Thermal Processing of Semiconductors—RTP '98*, edited by T. Hori, B. Lojek, Y. Tanabe, and R. P. S. Thakur, 72. Round Rock, TX: RTP Conference, 1998.
272. Das, J. H., Y. B. Jia, J. Y. Choi, R. P. S. Thakur, A. D. Daniel, A. J. Atanos, and S. P. Tay. In *Rapid Thermal and Other Short-Time Processing Technologies II*, edited by D.-L. Kwong, K. G. Reid, M. C. Öztürk, P. J. Timans, and F. Roozeboom, 147. Pennington, NJ: The Electrochemical Society, 2001.
273. Lee, M. H., C.-Y. Yu, F. Yuan, K.-F. Chen, C.-C. Lai, and W. Liu. *IEEE Trans. Semicond. Manuf.* 16 (2003): 656.
274. Arakawa, T., H. Fukuda, and S. Ohno. *IEEE Electron Device Lett.* 12 (1991): 66.
275. Weinberg, Z. A., D. R. Young, J. A. Calise, S. A. Cohen, J. C. DeLuca, and R. Deline. *Appl. Phys. Lett.* 45 (1984): 1204.
276. Baumvol, I. J. R., F. Stedile, J.-J. Ganem, I. Trimaille, and S. Rigo. *J. Electrochem. Soc.* 143 (1996): 1426.
277. Nulman, J., and J. P. Krusius. *Appl. Phys. Lett.* 47 (1985): 150.
278. Chang, C. C., A. Kamgar, and D. Kahng. *IEEE Electron Device Lett.* EDL-6 (1985): 476.
279. Hori, T., H. Iwasaki, Y. Naito, and H. Esaki. *IEEE Trans. Electron Devices* ED-34 (1987): 2238.
280. Hwang, H., W. Ting, D.-L. Kwong, and J. Lee. *IEEE Electron Device Lett.* 12 (1991): 495.
281. Fukuda, H., M. Yasuda, T. Iwabuchi, and S. Ohno. *IEEE Electron Device Lett.* 12 (1991): 587.
282. Hwang, H., W. Ting, D.-L. Kwong, and J. Lee. *Appl. Phys. Lett.* 59 (1991): 1581.
283. Bhat, M., D. J. Wristers, L.-K. Han, J. Yan, H. J. Fulford, and D.-L. Kwong. *IEEE Trans. Electron Devices* 42 (1995): 907.
284. Joshi, A. B., G. Q. Lo, D. K. Shih, and D. L. Kwong. *Proc. SPIE* 1393 (1990): 122.
285. Hori, T., and H. Iwasaki. *IEEE Electron Device Lett.* 9 (1988): 168.
286. Hori, T., H. Iwasaki, and K. Tsuji. *IEEE Trans. Electron Devices* 36 (1989): 340.
287. Lucovsky, G., C. R. Parker, Y. Wu, and J. R. Hauser. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 187.
288. Chen, C. H., Y. K. Fang, C. W. Yang, S. F. Ting, Y. S. Tsair, M. C. Yu, T. H. Hou., et al. *IEEE Electron Device Lett.* 22 (2001): 378.
289. Perera, R., A. Ikeda, R. Hattori, and Y. Kuroki. *Microelectron. Eng.* 65 (2003): 357.
290. Hegedus, A., C. S. Olsen, N. Kuan, and J. Madok. *IEEE Trans. Semicond. Manuf.* 16 (2003): 165.
291. Chung, H. Y. A., J. Niess, W. Dietl, G. Roters, W. Lerch, Z. Nenyei, A. Ludsteck., et al. *Semicond. Int.* 27 (2004): 73.
292. Ludsteck, A., W. Dietl, H. Y. Chung, C. Tolksdorf, J. Schulze, Z. Nenyei, and I. Eisele. *Mater. Res. Soc. Symp. Proc.* 786 (2004): E3.14.1.
293. Ludsteck, A., J. Schulze, I. Eisele, W. Dietl, H. Chung, Z. Nenyei, A. Bergmaier, and G. Dollinger. *J. Electrochem. Soc.* 152 (2005): G334.
294. Kim, Y. H., S. C. Song, H. F. Luan, J. C. Gelpey, A. Kepton, S. Levy, R. Bloom, and D.-L. Kwong. In *Rapid Thermal and Other Short-Time Processing Technologies*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, K. Reid, and D.-L. Kwong, 239. Pennington, NJ: The Electrochemical Society, 2000.
295. Shriver, M. A., T. K. Higman, S. A. Campbell, C. J. Taylor, and J. Roberts. *Mater. Res. Soc. Symp. Proc.* 611 (2000): C4.4.1.
296. Matsushita, D., K. Muraoka, Y. Nakasaki, K. Kato, S. Inumiya, K. Eguchi, and M. Takayanagi. In *2004 Symposium on VLSI Technology Technical Digest*, 172. Piscataway, NJ: IEEE, 2004.
297. Thakur, R. P. S., R. Hawthorne, V. K. Mathews, P. C. Fazan, C. Werkhoven, E. Granneman, and R. Wilhelm. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 195.
298. Ando, K., A. Ishitani, and K. Hamano. *Appl. Phys. Lett.* 59 (1991): 1081.

299. Shriver, M. A., A. M. Gabrys, T. K. Higman, and A. Campbell. *Mater. Res. Soc. Symp. Proc.* 670 (2001): K2.3.1.
300. Shi, X., M. Shriver, Z. Zhang, T. Higman, and S. A. Campbell. *J. Vac. Sci. Technol. A* 22 (2004): 1146.
301. Olsen, C. S., F. Nouri, M. Rubin, O. Laparra, and G. Scott. *Proc. SPIE* 3881 (1999): 215.
302. Mukhopadhyay, M., R. Rajivakshan, and G. Yong Lin Lee. In *IEEE International Symposium on Semiconductor Manufacturing, 2003*, 255. Piscataway, NJ: IEEE, 2003.
303. Wang, X., J. Liu, F. Zhu, N. Yamada, and L. Kwong. *IEEE Trans. Electron Devices* 51 (2004): 1798.
304. Conley, J. F. Jr., D. J. Tweet, Y. Ono, and G. Stecker. *Mater. Res. Soc. Symp.* 811 (2004): D1.3.1.
305. Green, M. L., T. Conard, B. Brijs, M.-Y. Ho, G. D. Wilk, P. I. Räisänen, T. Sorsch, and W. Vandervorst. In *Rapid Thermal and Other Short-Time Processing Technologies III*, edited by P. J. Timans, E. Gusev, F. Roozeboom, M. C. Öztürk, and D.-L. Kwong, 177. Pennington, NJ: The Electrochemical Society, 2002.
306. Zhan, N., K. L. Ng, H. Wong, M. C. Poon, and C. W. Kok. In *2003 IEEE Conference on Electron Devices and Solid-State Circuits*, 431. Piscataway, NJ: IEEE, 2003.
307. Lysaght, P., B. Foran, S. Stemmer, G. Bersuker, J. Bennett, R. Tichy, L. Larson, and H. R. Huff. *Microelectron. Eng.* 69 (2003): 182.
308. Takeuchi, H., and T.-J. King. *Mater. Res. Soc. Symp.* 811 (2004): D7.6.1.
309. Sun, S. C., and T.-F. Chen. *IEEE Electron Device Lett.* 17 (1996): 355.
310. Han, L. K., G. W. Yoon, D. L. Kwong, V. K. Matthews, and P. C. Fazan. *IEEE Electron Device Lett.* 15 (1994): 280.
311. Akbar, M. S., H.-J. Cho, R. Choi, C. S. Kang, C. Y. Kang, C. H. Choi, S. J. Rhee, Y. H. Kim, and J. C. Lee. *IEEE Electron Device Lett.* 25 (2004): 465.
312. Sekine, K., S. Inumiya, M. Sato, A. Kaneko, K. Eguchi, and Y. Tsunashima. In *2003 International Electron Devices Meeting Technical Digest*, 103. Piscataway, NJ: IEEE, 2003.
313. Quevedo-Lopez, M. A., M. R. Visokay, J. J. Chambers, M. J. Bevan, A. LiFatou, L. Colombo, M. J. Kim, B. E. Gnade, and R. M. Wallace. *J. Appl. Phys.* 97 (2005): 43508.
314. Lee, C., J. Choi, M. Cho, J. Park, C. S. Hwang, H. J. Kim, and J. Jeong. *J. Vac. Sci. Technol. B* 22 (2004): 1838.
315. Wang, J. C., Y. P. Hung, C. L. Lee, and F. Lei. *J. Electrochem. Soc.* 151 (2004): F17.
316. Bastos, P., R. P. Pezzi, L. Miotti, G. V. Soares, C. Driemeier, J. Morais, I. J. R. Baumvol, C. Hinkle, and G. Lucovsky. *Appl. Phys. Lett.* 84 (2004): 97.
317. Rubloff, G. W., and D. T. Bordonaro. *IBM J. Res. Dev.* 36, no. 2 (1992): 233.
318. Grant, J. M. *Mater. Res. Soc. Symp. Proc.* 387 (1995): 175.
319. Ma, Y., and M. L. Green. In *Advances in Rapid Thermal and Integrated Processing*, edited by F. Roozeboom, 217. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1996.
320. Glowacki, F., B. Froeschle, L. Deutschmann, I. Sagnes, D. Laviale, D. Bensahel, A. Halimaoui, F. Martin, and J. Bauer. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 229.
321. Froeschle, B., N. Sacher, F. Glowacki, T. Pompl, G. Innertsberger, and A. Gschwandtner. In *6th International Conference on Advanced Thermal Processing of Semiconductors—RTP '98*, edited by T. Hori, B. Lojek, Y. Tanabe, and R. P. S. Thakur, 90. Round Rock, TX: RTP Conference, 1998.
322. Frystack, D. C., J. Kuehne, R. Wise, B. Fowler, P. Grothe, J. Barnett, and G. Miner. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 221.
323. Thakur, R. P. S., S. J. DeBoer, E.-X. Ping, and A. Jesse. *IEEE Trans. Electron Devices* 45 (1988): 609.
324. Thakur, R. P. S., R. Hawthorne, V. K. Mathews, P. C. Fazan, C. J. Werkhoven, E. Granneman, and R. Wilhelm. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 195.
325. Xu, X., R. T. Kuehn, M. C. Öztürk, J. Wortman, R. J. Nemanich, G. S. Harris, and M. Maher. *J. Electron. Mater.* 22 (1993): 335.
326. Shih, H. H., J. Y. Wu, and W. Lur. In *6th International Conference on Advanced Processing of Semiconductors—RTP '98*, edited by T. Hori, B. Lojek, Y. Tanabe, and R. P. S. Thakur, 86. Round Rock, TX: RTP Conference, 1998.
327. Moslehi, M. M. *Proc. SPIE* 1393 (1990): 90.
328. Hill, C. *Mater. Res. Soc. Symp. Proc.* 1 (1981): 361.
329. Hill, C. *Mater. Res. Soc. Symp. Proc.* 13 (1983): 381.

330. Fair, R. B. In *Rapid Thermal Processing: Science and Technology*, edited by R. B. Fair, 169. San Diego, CA: Academic Press, 1993.
331. Jones, E. C., and E. Ishida. *Mater. Sci. Eng. Rep.* R24, no. 1–2 (1998): 1–80.
332. Timans, P. J., W. Lerch, J. Niess, S. Paul, N. Acharya, and Z. Nenyei. In *11th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2003*, edited by J. Gelpey, B. Lojek, Z. Nenyei, and R. Singh, 17. Piscataway, NJ: IEEE, 2003.
333. Hwang, J., H. Kennel, P. Packan, M. Taylor, M. Liu, R. James, and M. Kuhn. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices*, edited by F. Roozeboom, E. P. Gusev, L. J. Chen, M. C. Öztürk, D.-L. Kwong, and P. J. Timans, 35. Pennington, NJ: The Electrochemical Society, 2003.
334. Timans, P. J., W. Lerch, S. Paul, J. Niess, T. Huelsmann, and P. Schmid. *Solid State Technol.* 47 (2004): 35.
335. Ohuchi, K., K. Adachi, A. Hokazono, and Y. Toyoshima. *Mater. Res. Soc. Symp. Proc.* 717 (2002): C2.2.1.
336. Michel, A. E., W. Rausch, P. A. Ronsheim, and H. Kastl. *Appl. Phys. Lett.* 50 (1987): 416.
337. Stolk, P. A., H.-J. Gossmann, D. J. Eaglesham, D. C. Jacobson, C. S. Rafferty, G. H. Gilmer, M. Jaraíz, J. M. Poate, H. S. Luftman, and E. Haynes. *J. Appl. Phys.* 81 (1997): 6031.
338. Cristiano, F., B. Colombeau, B. de Mauduit, C. Bonafos, G. Benassayag, and A. Claverie. *Mater. Res. Soc. Symp. Proc.* 717 (2002): C5.7.1.
339. Pichler, P., and D. Stiebel. *Nucl. Instrum. Methods Phys. Res. B* 186 (2002): 256.
340. Stolk, P. A., F. N. Cubaynes, V. M. H. Meyssen, G. Mannino, N. E. B. Cowern, J. P. van Zijl, F. Roozeboom., et al. *Mater. Res. Soc. Symp. Proc.* 610 (2000): B3.1.1.
341. Cowern, N. E. B., G. Mannino, F. Roozeboom, P. A. Stolk, H. G. A. Huizing, J. G. M. van Berkum, and N. N. Toan. In *Advances in Rapid Thermal Processing*, edited by J. C. Gelpey, M. C. Öztürk, and J. Nakos, 125. Pennington, NJ: The Electrochemical Society, 1999.
342. Chakravarthi, S., A. H. Gencer, S. T. Dunham, and D. F. Downey. *Mater. Res. Soc. Symp. Proc.* 610 (2000): B4.8.1.
343. Lerch, W., B. Bayha, D. F. Downey, and E. Arevalo. In *Rapid Thermal and Other Short-Time Processing Technologies II*, edited by D.-L. Kwong, K. G. Reid, M. C. Öztürk, P. J. Timans, and F. Roozeboom, 321. Pennington, NJ: The Electrochemical Society, 2001.
344. Shishiguchi, S., A. Mineji, T. Hayashi, and S. Saito. In *1997 Symposium on VLSI Technology Digest of Technical Papers*, 89, Piscataway, NJ: IEEE, 1997.
345. Fiory, A. T., and K. K. Bourdelle. *Appl. Phys. Lett.* 74 (1999): 2658.
346. Mokhberi, A., P. B. Griffin, J. D. Plummer, E. Paton, S. McCoy, and K. Elliott. *IEEE Trans. Electron Devices* 49 (2002): 1183.
347. Fiory, A. T., K. K. Bourdelle, M. E. Lefrancois, D. M. Camm, and A. Agarwal. In *Advances in Rapid Thermal Processing*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, and J. Nakos, 133. Pennington, NJ: The Electrochemical Society, 1999.
348. Jain, A. *Mater. Res. Soc. Symp.* 810 (2004): C5.6.1.
349. Matsuda, T., S. Shishiguchi, and H. Kitajima. *Jpn. J. Appl. Phys.* 41 (2002): 451.
350. Paul, S., W. Lerch, X. Hebras, N. Cherkashin, and F. Cristiano. *Mater. Res. Soc. Symp. Proc.* 810 (2004): C5.4.1.
351. Bayha, B., D. Loeffelmacher, W. Lerch, D. F. Downey, and E. Arevalo. In *2000 Conference on Ion Implantation Technology*, 623. Piscataway, NJ: IEEE, 2000.
352. Bourdelle, K. K., A. T. Fiory, H.-J. L. Gossmann, and S. P. McCoy. *Mater. Res. Soc. Symp. Proc.* 610 (2000): J8.1.1.
353. Kim, H.-S., J.-H. Ahn, D.-M. Lee, K.-D. Yoo, S.-C. Lee, and K.-P. Suh. *J. Appl. Phys. 1* 39 (2000): 2172.
354. Son, J.-H., S.-H. Lee, J.-S. Lee, and Y. Lee. *Solid-State Electron.* 45 (2001): 7.
355. Sedgwick, T. O., A. E. Michel, V. R. Deline, S. A. Cohen, and J. B. Lasky. *J. Appl. Phys.* 63 (1988): 1452.
356. Mansoori, M. M., A. Jain, D. E. Mercer, L. Robertson, and P. Kohli. In *Rapid Thermal and Other Short-Time Processing Technologies III*, edited by P. J. Timans, E. Gusev, F. Roozeboom, M. C. Öztürk, and D.-L. Kwong, 389. Pennington, NJ: The Electrochemical Society, 2002.

357. Downey, D. F., J. W. Chow, W. Lerch, J. Niess, and S. D. Marcus. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 263.
358. Chow, J. W., and D. F. Downey. In *Rapid Thermal Processing '98*, edited by T. Hori, B. Lojek, Y. Tanabe, and R. P. S. Thakur, 105. Round Rock, TX: RTP '98, 1998.
359. Downey, D. F., S. Daryanani, M. Meloni, K. M. Brown, S. B. Felch, B. S. Lee, S. D. Marcus, and J. Gelpey. *Mater. Res. Soc. Symp. Proc.* 470 (1997): 299.
360. Srinivisa, R., V. Agarwal, J. Liu, D. F. Downey, and S. Banerjee. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 257.
361. Naem, A. A., A. R. Boothroyd, and I. D. Calder. *Mater. Res. Soc. Symp. Proc.* 23 (1984): 229.
362. Gelpey, J. C., K. Elliott, D. Camm, S. McCoy, J. Ross, D. F. Downey, and E. A. Arevalo. In *Rapid Thermal and Other Short-Time Processing Technologies III*, edited by P. J. Timans, E. Gusev, F. Roozeboom, M. C. Öztürk, and D.-L. Kwong, 313. Pennington, NJ: The Electrochemical Society, 2002.
363. McCoy, S. P., E. A. Arevalo, J. C. Gelpey, and D. F. Downey. In *12th International Conference on Advanced Thermal Processing of Semiconductors—RTP 2004*, edited by J. Gelpey, B. Lojek, Z. Nenyei, and R. Singh, 99. Piscataway, NJ: IEEE, 2004.
364. Satta, A., R. Lindsay, S. Severi, K. Henson, K. Maex, S. McCoy, J. Gelpey, and K. Elliott. *Mater. Res. Soc. Symp. Proc.* 810 (2004): 15.
365. Jain, S. H., P. B. Griffin, J. D. Plummer, S. McCoy, J. Gelpey, T. Selinger, and D. F. Downey. *J. Appl. Phys.* 96 (2004): 7357.
366. Suguro, K., T. Ito, K. Nishinohara, K. Matsuo, T. Iinuma, H. Itokawa, and Y. Kawase. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices II*, edited by M. C. Öztürk, E. P. Gusev, L. J. Chen, D.-L. Kwong, P. J. Timans, G. Miner, and F. Roozeboom, 39. Pennington, NJ: The Electrochemical Society, 2004.
367. Talwar, S., D. Markle, and M. Thompson. *Solid State Technol.* 46 (2003): 83.
368. Timans, P. J., and N. Acharya. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices II*, edited by M. C. Öztürk, E. P. Gusev, L. J. Chen, D.-L. Kwong, P. J. Timans, G. Miner, and F. Roozeboom, 11. Pennington, NJ: The Electrochemical Society, 2004.
369. Timans, P. J., and N. Acharya. *Semiconductor Fabtech*. 22nd ed., 83, 2004.
370. Thompson, K., J. H. Booske, R. L. Ives, J. Lohr, Y. A. Gorelov, and K. Kajiwara. *Mater. Res. Symp. Proc.* 810 (2004): C5.3.1.
371. Ito, T., K. Matsuo, H. Itokawa, T. Itani, N. Tamaoki, Y. Honguh, K. Suguro, et al. In *Extended Abstracts of the 5th International Workshop on Junction Technology 2005*, S4-3, Piscataway, NJ: IEEE, 2005.
372. Shima, A., H. Ashihara, T. Mine, Y. Goto, M. Horiuchi, Y. Wang, S. Talwar, and A. Hiraiwa. In *2003 International Electron Devices Meeting Technical Digest*, 20.4.1. Piscataway, NJ: IEEE, 2003.
373. Shima, A., H. Ashihara, A. Hiraiwa, T. Mine, and Y. Goto. *IEEE Trans. Electron Devices* 52 (2005): 1165.
374. Earles, S., M. Law, R. Brindos, K. Jones, S. Talwar, and S. Corcoran. *IEEE Trans. Electron Devices* 49 (2002): 1118.
375. Lindsay, R., B. J. Pawlak, P. Stolk, and K. Maex. *Mater. Res. Soc. Symp. Proc.* 717 (2002): C2.1.1.
376. Lerch, W., S. Paul, J. Niess, F. Cristiano, Y. Lamrani, P. Calvo, N. Cherkashin, D. F. Downey, and E. A. Arevalo. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices II*, edited by M. C. Öztürk, E. P. Gusev, L. J. Chen, D.-L. Kwong, P. J. Timans, G. Miner, and F. Roozeboom, 90. Pennington, NJ: The Electrochemical Society, 2004.
377. Suzuki, K., H. Tashiro, K. Narita, and Y. Kataoka. *IEEE Trans. Electron Devices* 51 (2004): 663.
378. Tsuji, K., K. Takeuchi, and T. Mogami. In *Symposium on VLSI Technology, 1999, Digest of Technical Papers*, 9–10. Piscataway, NJ: IEEE, 1999.
379. Landi, E., S. Guimaraes, and S. Solmi. *Appl. Phys. A* 44 (1987): 135.
380. Pawlak, B. J., W. Vandervorst, R. Lindsay, I. De Wolf, F. Roozeboom, R. Delhougne, A. Benedetti., et al. *Mater. Res. Soc. Symp. Proc.* 810 (2004): C9.6.1.
381. Lauwers, A., R. Lindsay, K. Henson, S. Severi, A. Akheyar, B. J. Pawlak, M. de Potter, and K. Maex. *Mater. Res. Soc. Symp. Proc.* 810 (2004): C2.2.1.

382. Borland, J. O. *Mater. Res. Soc. Symp. Proc.* 717 (2002): C1.1.1.
383. Timans, P. J., N. Acharya, and I. Amarilio. In *Rapid Thermal and Other Short-Time Processing Technologies*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, K. Reid, and D.-L. Kwong, 375. Pennington, NJ: The Electrochemical Society, 2000.
384. Fiory, A. T., and K. Bourdelle. *Mater. Res. Soc. Symp. Proc.* 610 (2000): B3.3.1.
385. Fiory, A. T., K. K. Bourdelle, Y. Chen, Y. Ma, J. M. McKinley, P. K. Roy, and H. W. Koh. In *Rapid Thermal and Other Short-Time Processing Technologies II*, edited by D.-L. Kwong, K. G. Reid, M. C. Öztürk, P. J. Timans, and F. Roozeboom, 89. Pennington, NJ: The Electrochemical Society, 2001.
386. Cubaynes, F. N., P. A. Stolk, J. Verhoeven, F. Roozeboom, and P. H. Woerlee. *Mater. Sci. Semicond. Process.* 4 (2001): 351.
387. Roh, K., S. Youn, S. Yang, and Y. Roh. *J. Vac. Sci. Technol. A* 19 (2001): 1562.
388. Ma, Y., K. Z. Ahmed, K. L. Cunningham, C. S. Olsen, T. Y. B. Leung, R. C. McIntosh, A. J. Mayur., et al. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices II*, edited by M. C. Öztürk, E. P. Gusev, L. J. Chen, D.-L. Kwong, P. J. Timans, G. Miner, and F. Roozeboom, 230. Pennington, NJ: The Electrochemical Society, 2004.
389. Chong, Y. F., H.-J. L. Gossmann, K.-L. Pey, M. O. Thompson, A. T. S. Wee, and C. H. Tung. *IEEE Trans. Electron Devices* 51 (2004): 669.
390. Osburn, C. M. In *Rapid Thermal Processing*, edited by R. B. Fair, 227. San Diego, CA: Academic Press, 1993.
391. van den Hove, L., and D. F. Keersmaecker. In *Rapid Thermal Processing for VLSI*, edited by R. A. Levy, 53. New York: Plenum Press, 1989.
392. Lasky, J. B., J. S. Nakos, O. J. Cain, and J. Geiss. *IEEE Trans. Electron Devices* 38 (1991): 262.
393. Kittl, J. A., D. A. Prinslow, G. Misium, and F. Pas. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 175.
394. Schreutelkamp, R. J., P. Vandenabeele, B. Deweerdt, R. Verbeeck, and K. Maex. *Appl. Surf. Sci.* 73 (1993): 162.
395. Besser, P. R., S. Chan, E. Paton, T. Kammler, D. Brown, P. King, and L. Pressley. *Mater. Res. Soc. Symp. Proc.* 766 (2003): E10.1.1.
396. Kittl, J. A., A. Lauwers, O. Charmirian, M. van Dal, A. Akheyar, O. Richard, J. G. Lisoni, M. De Potter, R. Lindsay, and K. Maex. *Mater. Res. Soc. Symp. Proc.* 765 (2003): D7.5.1.
397. Kittl, J. A., A. Lauwers, O. Chamirian, M. A. Pawlak, M. van Dal, A. Akheyar, M. De Potter., et al. *Mater. Res. Soc. Symp.* 810 (2004): C2.1.1.
398. Nanda, A. K., S. Meester, and W. Wilkins. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 111.
399. Shenai, K. *IEEE Trans. Semicond. Manuf.* 4 (1991): 1.
400. Jones, R. E., and T. C. Mele. *IEEE Trans. Semicond. Manuf.* 4 (1991): 281.
401. Lu, C.-Y., J. J. Sung, R. Liu, N.-S. Tsai, R. Singh, S. J. Hillenius, and H. C. Kirsch. *IEEE Trans. Electron Devices* 38 (1991): 246.
402. Brat, T., C. M. Osburn, T. Finstad, J. Liu, and B. Ellington. *J. Electrochem. Soc.* 133 (1986): 1451.
403. Lin, X. W., and D. Pramanik. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 181.
404. Hu, Y. Z., S. P. Tay, J. Yang, R. Thakur, P. M. Smith, and G. Bailey. In *Advances in Rapid Thermal Processing*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, and J. Nakos, 229. Pennington, NJ: The Electrochemical Society, 1999.
405. Ganapathiraman, R., S. Koh, Z. Ma, L. H. Allen, and S. Lee. *Mater. Res. Soc. Symp. Proc.* 303 (1993): 63.
406. Kappius, L., and R. T. Tung. In *Rapid Thermal and Other Short-Time Processing Technologies*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, K. Reid, and D.-L. Kwong, 139. Pennington, NJ: The Electrochemical Society, 2000.
407. Ren, L. P., P. Liu, G. Z. Pan, and J. Woo. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 313.
408. Kittl, J. A., Q. Z. Hong, H. Yang, N. Yu, M. Rodder, P. P. Apte, W. T. Shiau, C. P. Chao, T. Breedjik, and F. Pas. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 331.
409. Li, H., G. Vereecke, K. Maex, and L. Froyen. *J. Electrochem. Soc.* 148 (2001): G344.
410. Wang, Q. F., A. Lauwers, B. Deweerdt, R. Verbeeck, F. Loosen, and K. Maex. *IEEE Trans. Semicond. Manuf.* 8 (1995): 449.
411. Detavernier, C., R. L. van Meirhaeghe, and K. Maex. *Mater. Res. Soc. Symp. Proc.* 670 (2001): K7.4.1.

412. Sohn, D. K., J. Park, B. H. Lee, J. Bae, K. S. Oh, S. K. Lee, J. S. Byun, and J. J. Kim. In *IEDM 1998 Proceedings*, 1005. Piscataway, NJ: IEEE, 1998.
413. Chen, Y., M. W. Lippitt, H. Chew, and M. M. Moller. *IEEE Trans. Electron Devices* 50 (2003): 2120.
414. Chen, W.-M., S. Pozder, Y. Limb, A. R. Sitaram, and B. Fiordalice. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 163.
415. Wacquant, F., C. Regnier, M.-T. Basso, C. Julien, A. Humbert, and C. Jenny. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices*, edited by F. Roozeboom, E. P. Gusev, L. J. Chen, M. C. Öztürk, D.-L. Kwong, and P. J. Timans, 191. Pennington, NJ: The Electrochemical Society, 2003.
416. Buschbaum, S., O. Fursenko, D. Bolze, D. Wolansky, V. Melnik, J. Niess, and W. Lerch. *Microelectron. Eng.* 76 (2004): 311.
417. Chen, K. M., H. J. Huang, C. Y. Chang, T. Y. Huang, G. W. Huang, and L. P. Chen. *Mater. Chem. Phys.* 69 (2001): 84.
418. Lavoie, C., R. Purtell, C. Coïa, C. Detavernier, P. Desjardins, J. Jordan-Sweet, C. Cabral Jr., F. M. d'Heurle, and J. M. E. Harper. In *Rapid Thermal and Other Short-Time Processing Technologies III*, edited by P. J. Timans, E. Gusev, F. Roozeboom, M. C. Öztürk, and D.-L. Kwong, 455. Pennington, NJ: The Electrochemical Society, 2002.
419. Coïa, C., C. Lavoie, F. M. d'Heurle, C. Detavernier, P. Desjardins, and A. J. Kellock. In *Advanced Gate Stack, Source/Drain and Channel Engineering for Si-Based CMOS: New Materials, Processes, and Equipment*, edited by E. P. Gusev, L. J. Chen, H. Iwai, D.-L. Kwong, M. C. Öztürk, F. Roozeboom, and P. J. Timans, 585. Pennington, NJ: The Electrochemical Society, 2005.
420. Lauwers, A., J. A. Kittl, M. van Dal, O. Chamirian, R. Lindsay, M. de Potter, C. Demeurisse., et al. *Microelectron. Eng.* 76 (2004): 303.
421. Hou, T.-H., T.-F. Lei, and T.-S. Chao. *IEEE Electron Device Lett.* 20 (1999): 572.
422. Niess, J., S. Paul, S. Buschbaum, P. Schmid, and W. Lerch. *Mater. Sci. Eng. B* 114–115 (2004): 141.
423. Zhao, F. F., Z. X. Shen, J. Z. Zheng, W. Z. Gao, T. Osipowicz, C. H. Pang, P. S. Lee, and A. K. See. *Mater. Res. Soc. Symp. Proc.* 716 (2002): B1.8.1.
424. Ma, D., D. Z. Chi, W. D. Wang, A. S. W. Wong, and S. J. Chua. *Mater. Res. Soc. Symp. Proc.* 716 (2002): B1.9.1.
425. Öztürk, M. C., J. Liu, and H. Mo. In *2003 International Electron Devices Meeting Technical Digest*, 497. Piscataway, NJ: IEEE, 2003.
426. Borisenko, V. E., and P. J. Hesketh. *Rapid Thermal Processing of Semiconductors*. 149. New York: Plenum Press, 1997.
427. Bevk, J., M. Furtsch, G. E. Georgiou, S. J. Hillenius, D. Schielein, T. Schiml, P. J. Silverman, and H. S. Luftman. *Mater. Res. Soc. Symp. Proc.* 429 (1995): 115.
428. Blosse, A. P. *Mater. Res. Soc. Symp. Proc.* 525 (1998): 371.
429. Yamashita, T., Y. Nishida, K. Hayashi, T. Eimori, M. Inuishi, and Y. Ohji. *Jpn. J. Appl. Phys. 1* 43 (2004): 1799.
430. Katiyar, M., G. S. Samal, R. K. Gupta, Deepak, P. K. Sahoo, V. N. Kulkarni, and O. Adetutu. *Mater. Res. Soc. Symp. Proc.* 670 (2001): K5.7.1.
431. Byun, J. S., B. H. Lee, J.-S. Pork, D.-K. Sohn, S. J. Choi, and J. J. Kim. *J. Electrochem. Soc.* 145 (1998): 3228.
432. Cho, W.-J., and S. Lee. *Jpn. J. Appl. Phys. 1* 42 (2003): 2615.
433. Rao, V., J. Morgan, W. Hoesler, J. Barden, Y. Karzhavin, P. van Holt, R. Petter, H. Ollendorf, K. Christensen, and D. Ricks. In *2000 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, 340. Piscataway, NJ: IEEE, 2000.
434. Jin, S., H. Bender, R. A. Donaton, and K. Maex. *J. Mater. Res.* 14 (1999): 2577.
435. Larrieu, G., E. Dubois, and X. Wallart. *Mater. Res. Soc. Symp. Proc.* 765 (2003): D7.9.1.
436. Pant, A. K., S. P. Murarka, C. Shepard, and W. Lanford. *J. Appl. Phys.* 72 (1992): 1833.
437. Jang, M., Y. Kim, J. Shin, and S. Lee. *IEEE Electron Device Lett.* 26 (2005): 354.
438. Li, M.-F., S. Lee, S. Zhu, R. Li, J. Chen, A. Chin, and D. L. Kwong. In *Advanced Gate Stack, Source/Drain and Channel Engineering for Si-Based CMOS: New Materials, Processes, and Equipment*, edited by E. P. Gusev, L. J. Chen, H. Iwai, D.-L. Kwong, M. C. Öztürk, F. Roozeboom, and P. J. Timans, 301. Pennington, NJ: The Electrochemical Society, 2005.

439. Lee, C.-J., and Y.-K. Sung. *J. Electron. Mater.* 22 (1993): 717.
440. Perez-Rigueiro, J., C. Jiminez, L. Vasquez, R. Perez-Casero, and J. M. Martinez-Duart. *Surf. Coat. Technol.* 80 (1996): 72.
441. Farahani, M. M., S. Garg, and B. T. Moore. *J. Electrochem. Soc.* 141 (1994): 479.
442. Cohen, B., and J. Nulman. *Mater. Res. Soc. Symp. Proc.* 92 (1987): 171.
443. Kermani, A., and J. Kuehne. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 241.
444. Yun, E. J., H. G. Chun, K. Jung, D. L. Kwong, and S. Lee. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 255.
445. Yao, G. D., Y. C. Lu, S. Prasad, W. Hata, F. S. Chen, and H. Zhang. *Mater. Res. Soc. Symp. Proc.* 303 (1993): 103.
446. Lee, C. Y., H. Yen, S. T. Hsia, D. Liu, N. Shah, K. Feldmeier, and Y. Wasserman. *Mater. Res. Soc. Symp. Proc.* 387 (1995): 383.
447. Drynan, J. M., and K. Koyama. *Mater. Res. Soc. Symp. Proc.* 387 (1995): 419.
448. Drynan, J. M., and K. Koyama. *Mater. Res. Soc. Symp. Proc.* 429 (1996): 141.
449. Lee, S. S., C. S. Galovich, K. P. Fuchs, D. L. Kwong, J. Hirvonen, and J. Huang. *Mater. Res. Soc. Symp. Proc.* 146 (1989): 217.
450. Herner, S. B., Y. Tanaka, H. Zhang, and G. Ghanayem. *J. Electrochem. Soc.* 147 (2000): 1982.
451. Hwang, C.-C., C.-C. Jaing, M.-J. Lai, J.-S. Chen, S. Huang, M.-H. Juang, and H.-C. Cheng. *Electrochem. Solid-State Lett.* 3 (2000): 563.
452. Abe, K., Y. Harada, M. Yoshimaru, and H. Onoda. *J. Vac. Sci. Technol. B* 22 (2004): 721.
453. Morgan, A. E., E. K. Broadbent, K. N. Ritz, D. K. Sadana, and J. Burrow. *J. Appl. Phys.* 64 (1988): 344.
454. Thakur, R. P. S., F. Gonzalez, R. Hawthorne, V. Ward, and N. Jeng. *Mater. Res. Soc. Symp. Proc.* 303 (1993): 283.
455. Maxim, M., M. Moinpour, J. Chu, H. Nguyen, P. Freiberger, and N. Stenton. *Mater. Res. Soc. Symp. Proc.* 342 (1994): 289.
456. Weimer, R. A. In *First International Symposium on ULSI Process Technology*, edited by G. Bronner, C. L. Claeys, and R. B. Fair, 59. Pennington, NJ: The Electrochemical Society, 1999.
457. Iyer, R., R. P. S. Thakur, H. Rhodes, R. Liao, R. Rosler, and E. Yieh. *J. Electrochem. Soc.* 143 (1996): 3366.
458. Das, J. H., Y. Brichko, A. D. Daniel, D. Clarke, K. Kapkin, and S. Al-Lami. In *7th International Conference on Advanced Processing of Semiconductors—RTP '99*, edited by H. Kitayama, B. Lojek, G. Miner, and A. Tillmann, 67. Round Rock, TX: RTP-Conference, 1999 (and also 181).
459. Shah, N., J. McVittie, N. Sharif, J. Nulman, and A. Gat. *Mater. Res. Soc. Symp. Proc.* 52 (1986): 233.
460. Lee, G. G., K. Fujihara, J. M. Ha, H. K. Kang, and M. Y. Lee. In *4th International Conference on Advanced Thermal Processing of Semiconductors—RTP '96*, edited by R. B. Fair, M. L. Green, B. Lojek, and R. P. S. Thakur, 30. Round Rock, TX: RTP '96, 1996.
461. Baker, F., A. Ballantine, E. Fisch, and W. Hodge. In *Advanced Semiconductor Manufacturing Conference and Workshop, 1999*, 394. Piscataway, NJ: IEEE, 1999.
462. Han, S. H., N. S. Kim, D. J. Son, M. Mukhopadhyay, W. Y. Wong, G. Zhang, and I. S. Goh. In *Proceedings 2004 Non-Volatile Memory Technology Symposium*, 70. Piscataway, NJ: IEEE, 2004.
463. Sharangpani, R., and S.-P. Tay. In *10th IEEE International Conference on Advanced Thermal Processing of Semiconductors—RTP 2002*, edited by J. Gelpey, B. Lojek, Z. Nenyei, and R. Singh, 143. Piscataway, NJ: IEEE, 2002.
464. Sharangpani, R., R. Singh, M. Drews, and K. Ivey. *J. Electron. Mater.* 26 (1997): 402.
465. Bremmer, J., D. Gray, Y. Liu, K. Gruszynski, and S. Marcus. *Mater. Res. Soc. Symp. Proc.* 565 (1999): 273.
466. Clarke, D., V. Bhaskaran, J. Sanchez, E. Broadbent, and R. Thakur. In *7th International Conference on Advanced Thermal Processing of Semiconductors—RTP '99*, edited by H. Kitayama, B. Lojek, G. Miner, and A. Tillmann, 113. Colorado Springs, CO: RTP '99, 1999.
467. Steinlesberger, G., M. Engelhardt, G. Schindler, W. Steinhogl, M. Traving, W. Honlein, and E. Bertagnolli. *Mater. Res. Soc. Symp. Proc.* 766 (2003): 379.
468. Kwon, D., H. Park, S. Ghosh, C. Lee, H. T. Jeon, and J. G. Lee. *J. Korean Phys. Soc. pt. 1* 44 (2004): 1108.

469. Oh, J., H. Lee, A. Paul, and C. Lee. *Jpn. J. Appl. Phys. 1* 40 (2001): 5294.
470. Jiang, Q.-T., A. Frank, R. H. Havemann, V. Parihar, and M. Nowell. In *Symposium on VLSI Technology. Digest of Technical Papers*, 139. Tokyo, Japan: The Japan Society of Applied Physics, 2001.
471. Beyer, G. P., P. Kitabjian, S. H. Brongersma, J. Proost, H. Bender, E. Richard, I. Vervoort, P. Hey, P. Zhang, and K. Maex. In *Advanced Metallization Conference 1999 (AMC 1999). Proceedings of the Conference*, edited by M. E. Gross, T. Gessner, N. Kobayashi, and Y. Yasuda, 167. Warrendale, PA: Materials Research Society, 2000.
472. Hu, Y. Z., R. Sharangpani, and S.-P. Tay. In *Rapid Thermal and Other Short-Time Processing Technologies*, edited by F. Roozeboom, J. C. Gelpey, M. C. Öztürk, K. Reid, and D.-L. Kwong, 329. Pennington, NJ: The Electrochemical Society, 2000.
473. Hu, Y. Z., R. Sharangpani, and S.-P. Tay. *J. Electrochem. Soc.* 148 (2001): G669.
474. Thompson, S. E. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices II*, edited by M. C. Öztürk, E. P. Gusev, L. J. Chen, D.-L. Kwong, P. J. Timans, G. Miner, and F. Roozeboom, 412. Pennington, NJ: The Electrochemical Society, 2004.
475. Kutsukake, K., N. Usami, K. Fujiwara, T. Ujihara, G. Sazaki, K. Nakajima, B. Zhang, and Y. Segawa. *Appl. Surf. Sci.* 224 (2004): 95.
476. Maiti, C. K., S. K. Samanta, S. Chatterjee, G. K. Dalapati, and L. K. Bera. *Solid-State Electron.* 48 (2004): 1369.
477. Xia, G., H. M. Nayfeh, M. L. Lee, E. A. Fitzgerald, D. A. Antoniadis, D. H. Anjum, J. Li, R. Hull, N. Klymko, and J. L. Hoyt. *IEEE Trans. Electron Devices* 51 (2004): 2136.
478. Lee, K. L., F. Cardone, P. Saunders, P. Kozlowski, P. Ronsheim, H. Zhu, J. Li, J. Chu, K. Chan, and M. Ieong. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices II*, edited by M. C. Öztürk, E. P. Gusev, L. J. Chen, D.-L. Kwong, P. J. Timans, G. Miner, and F. Roozeboom, 71. Pennington, NJ: The Electrochemical Society, 2004.
479. Rodriguez, J. A., A. Llobera, and C. Dominguez. *J. Mater. Sci. Lett.* 19 (2000): 1399.
480. Charavel, R., B. Olbrechts, and J.-P. Raskin. *Proc. SPIE* 5116 (2003): 596.
481. Guobing, Z., H. Yilong, T. Dayu, L. Shimei, W. Tiesong, and W. Guoying. *Chin. J. Semicond.* 20 (1999): 463.
482. Gilmer, D. C., C. Hobbs, J. Grant, R. Hegde, H. Tseng, D. Triyoso, D. Roan., et al. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices*, edited by F. Roozeboom, E. P. Gusev, L. J. Chen, M. C. Öztürk, D.-L. Kwong, and P. J. Timans, 345. Pennington, NJ: The Electrochemical Society, 2003.
483. Krivokapic, Z., and W. D. Heavlin. *IEEE Trans. Semicond. Manuf.* 15 (2002): 144.
484. Cabral, C., C. Lavoie, A. S. Ozcan, R. S. Amos, V. Narayanan, E. P. Gusev, J. L. Jordan-Sweet, and J. M. E. Harper. *J. Electrochem. Soc.* 151 (2004): F283.
485. Bae, S. H., W. P. Bai, H. C. Wen, S. Mathew, L. K. Bera, N. Balasubramanian, N. Yamada, M. F. Li, and D. L. Kwong. In *2004 Symposium on VLSI Technology Technical Digest*, 188. Piscataway, NJ: IEEE, 2004.
486. Li, T.-L., C.-H. Hu, W.-L. Ho, H. C.-H. Wang, and C.-Y. Chang. *IEEE Trans. Electron Devices* 52 (2005): 1172.
487. Westlinder, J., G. Sjoblom, and J. Olsson. *Microelectron. Eng.* 75 (2004): 389.
488. Maszara, W. P. In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices II*, edited by M. C. Öztürk, E. P. Gusev, L. J. Chen, D.-L. Kwong, P. J. Timans, G. Miner, and F. Roozeboom, 341. Pennington, NJ: The Electrochemical Society, 2004.
489. Cabral, C. Jr., J. Kedzierski, B. Linder, S. Zafar, V. Narayanan, S. Fang, A. Steegen, P. Kozlowski, R. Carrurthers, and R. Jammy. In *2004 Symposium on VLSI Technology Technical Digest*, 184. Piscataway, NJ: IEEE, 2004.
490. Kittl, J. A., A. Lauwers, M. A. Pawlak, C. Demeurisse, K. G. Anil, A. Veloso, M. J. H. van Dal., et al. In *Advanced Gate Stack, Source/Drain and Channel Engineering for Si-Based CMOS: New Materials, Processes, and Equipment*, edited by E. P. Gusev, L. J. Chen, H. Iwai, D.-L. Kwong, M. C. Öztürk, F. Roozeboom, and P. J. Timans, 225. Pennington, NJ: The Electrochemical Society, 2005.

491. Pawlak, M. A., J. A. Kittl, T. Janssens, A. Lauwers, W. Vandervorst, K. G. Anil, T. Schram., et al. In *Advanced Gate Stack, Source/Drain and Channel Engineering for Si-Based CMOS: New Materials, Processes, and Equipment*, edited by E. P. Gusev, L. J. Chen, H. Iwai, D.-L. Kwong, M. C. Öztürk, F. Roozeboom, and P. J. Timans, 241. Pennington, NJ: The Electrochemical Society, 2005.
492. Yu, D. S., C. H. Wu, C. H. Huang, A. Chin, W. J. Chen, Z. Chunxiang, M. F. Li, and D.-L. Kwong. *IEEE Electron Device Lett.* 24 (2003): 739.
493. van Dal, M. J. H., A. Lauwers, J. Cunniffe, R. Verbeeck, C. Vrancken, C. Demeurisse, T. Dao., et al. In *Advanced Gate Stack, Source/Drain and Channel Engineering for Si-Based CMOS: New Materials, Processes, and Equipment*, edited by E. P. Gusev, L. J. Chen, H. Iwai, D.-L. Kwong, M. C. Öztürk, F. Roozeboom, and P. J. Timans, 233. Pennington, NJ: The Electrochemical Society, 2005.
494. Matsuda, S., T. Sato, H. Yoshimura, Y. Takegawa, A. Sudo, I. Mizushima, Y. Tsunashima, and Y. Toyoshima. In *International Electron Devices Meeting Technical Digest*, 137. Piscataway, NJ: IEEE, 1998.
495. Choi, Y.-K., D. Ha, E. Snow, J. Bokor, and T.-J. King. In *International Electron Devices Meeting Technical Digest*, 177. Piscataway, NJ: IEEE, 2003.
496. Choi, Y.-K., N. Lindert, P. Xuan, S. Tang, D. Ha, E. Anderson, T.-J. King, J. Bokor, and C. Hu. In *2001 International Electron Devices Meeting Technical Digest*, 19.1.1. Piscataway, NJ: IEEE, 2001.
497. Xiong, W., G. Gebara, J. Zaman, M. Gostkowski, B. Nguyen, G. Smith, D. Lewis., et al. *IEEE Electron Device Lett.* 25 (2004): 541.
498. Chui, C. O., K. Gopalakrishnan, P. B. Griffin, J. D. Plummer, and K. C. Saraswat. *Appl. Phys. Lett.* 83 (2003): 3275.
499. Lee, K. Y., S. L. Liew, S. J. Chua, D. Z. Chi, H. P. Sun, and X. Q. Pan. *Mater. Res. Soc. Symp. Proc.* 810 (2004): C2.4.1.

12

Low-*k* Dielectrics

12.1 Introduction **12**-1
12.2 Channel Crack Failures **12**-3
Film Thickness Effects • Extracting Materials Properties
12.3 Elastic Constraint Effects **12**-9
12.4 Pattern Layout Effects **12**-13
12.5 Environmental Effects **12**-15
Interfacial Adhesion • Channel Cracking • Diffusion Studies
12.6 Conclusion **12**-21
References **12**-21

Ting Y. Tsui
Andrew J. McKerrow
Texas Instruments, Inc.

12.1 Introduction

Advances in manufacturing technology have provided the semiconductor industry with the means of producing products of increasing complexity. Scaling the minimum metal feature size that can be patterned robustly has provided a means increasing packing density in the front-end-of-the-line (FEOL) and the back-end-of-the-line (BEOL). While smaller FEOL features, i.e., gate length, typically exhibit improved performance, and it is possible to increase the density of such components that can be integrated per unit area of die, scaling in the BEOL can significantly degrade performance. The interconnect performance metric that is typically discussed in terms of adverse affects of geometric scaling across multiple technology generations is resistance–capacitance (RC) delay. This performance metric evaluates the combined effects of metal resistance and insulator capacitance on the propagation of a signals in the interconnect. In addition to design solutions aimed at minimizing interconnect RC delay, the semiconductor industry has investigated novel materials solutions to reduce the impact of resistance and capacitance on performance. From a metallization perspective, the wide-scale introduction of copper metallization has provided improvement in interconnect resistance, in comparison with older aluminum–copper metallization, and this topic will be covered elsewhere in this handbook. Efforts to reduce the effect of interconnect capacitance on RC delay have focused on the development, characterization, and integration of materials with a lower dielectric constant than silicon dioxide (SiO_2), the insulator most typically used in aluminum interconnects. These materials are most often referred to as low-dielectric constant or low-*k* materials and are the focus of this chapter.

The first edition of the Handbook of Semiconductor Manufacturing Technology included an overview of low-dielectric constant materials that were available at that time, characterization of their materials properties, and lessons learned from integrating these materials [1]. In the intervening period between editions of the Handbook, the semiconductor industry widely adopted fluorosilicate glass (FSG) as its first low-dielectric constant material for the 130-nm technology node. Typical FSG films were characterized by a dielectric constant (k) of 3.6–3.8, representing an approximate 10%–14% reduction

in dielectric constant relative to conventional BEOL SiO_2 ($k=4.2$). The integration of FSG with copper metallization provided numerous technical challenges, but in terms of materials properties, including mechanical strength, coefficient of thermal expansion, and thermal conductivity, FSGs were similar to the SiO_2 insulator that it replaced. At the 90-nm technology node, efforts to minimize the negative impacts of scaling on interconnect RC delay necessitated the wide-scale introduction of materials with a dielectric constant in the range of $k=2.8$–3.0. Initial industry efforts to develop such materials focused on two classes of materials: glasses deposited by plasma-enhanced chemical vapor deposition (Pe-CVD) or spin-on polymers. Over time the majority of semiconductor industry has adopted one class of low-k materials for the 90- and 65-nm technology nodes; Pe-CVD glasses commonly referred to as organosilicate glasses (OSGs) or carbon-doped oxides (CDOs).

Organosilicate glasses are routinely deposited in a parallel plate Pe-CVD reactor at temperatures between 350 and 400°C using a wide variety of organosilane precursors, including methyl silane, trimethylsilane, dimethyl dimethoxy silane, and tetramethyl cyclotetrasiloxane. The product of reaction of these precursors in a plasma is a glass that can be described as a silicon–oxygen network in which methyl (CH_3) groups are chemically bonded to silicon and thereby reduce the silicon–oxygen bond density, in comparison with conventional Pe-CVD SiO_2 [2,3]. Blanket films of these materials are characterized by dielectric constants ranging from 2.5 to 3.1 depending on deposition conditions [3]. The bonding structure of CH_3 groups in the glass can be well characterized by Fourier transform infrared absorption spectroscopy and a representative OSG absorption spectrum is shown in Figure 12.1. In addition to expected vibrational modes of silicon–oxygen bonds, this spectrum also absorptions that can be assigned to silicon–methyl, silicon–dimethyl, and silicon–hydrogen functionality in the film. Similar observations were reported by Grill et al. [4,5].

As noted above, the incorporation of methyl functionality in the glass comes at the expense of those silicon–oxygen bonds that typically result in the robust mechanical properties of SiO_2. Nanoindentation hardness, elastic modulus, dielectric constant, and density measurements for representative Pe-CVD OSG thin films are listed in Table 12.1 together with similar measurements for bulk-fused silica.

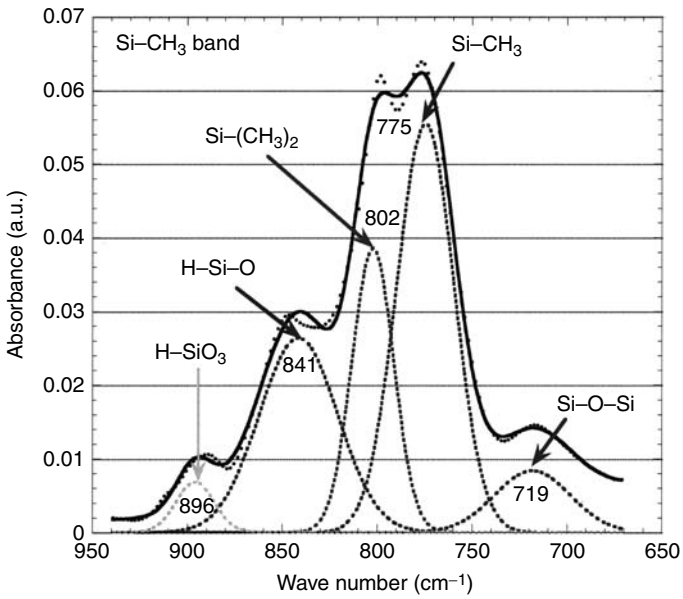


FIGURE 12.1 Fourier transform infrared absorption spectrum of commercially available organosilicate glass (OSG).

TABLE 12.1 Typical Nanoindentation Hardness, Elastic Modulus, Dielectric Constant of Fused Silica and Commercially Available Organosilicate Glass (OSG) Thin Films

	H (GPa)	*E* (GPa)	*K*	Density (g/cm^3)
Fused silica	10	72	3.78[a]	2.2[a]
OSG	1–2	3.0–16[b]	2.6–3.1[b]	1.27–1.44[b]

[a] Kingery, W. D., H. K. Bowen, and D. R. Uhlmann. *Introduction to Ceramics*, Wiley, New York, 1976.
[b] Tsui, T. Y., A. J. Griffin, Jr., R. Fields, J. M. Jacques, A. J. McKerrow, and J. J. Vlassak. *Thin Solid Films*, 515, 2257–2261, 2006.

The information included in this table shows that OSG films are approximately two times less dense and at least three times more compliant than bulk silica. These materials properties, coupled with residual tensile stress in OSG thin films, pose a significant risk of deformation and fracture for this class of low-*k* materials. As these failure modes are not similar to what was previously encountered in the semiconductor industry with more mechanically robust dielectrics, we will devote this chapter to reviewing common fracture mechanisms observed in the low-*k* materials and the techniques used to characterize them. This discussion is intended to reveal a range of mechanical properties and issues that exceed what can be adequately captured in nanoindentation modulus/hardness testing.

Looking beyond the 90- and 65-nm technology nodes, interconnect performance issues resulting from further scaling of minimum feature size is necessitating the development of materials characterized by a dielectric constant of $k = 2.5$ or less. These ultra-low-dielectric constant (ULK) materials are still being developed, but a common approach to preparing such films includes Pe-CVD deposition of an OSG film containing a small molecule that incorporates in the bulk of the film during deposition. A subsequent energetic cure (i.e., thermal, UV, e-beam) liberates the small molecule, or its fragments, leaving behind an OSG film with significant porosity. Given the similarity in chemical bonding in these ULK glasses with the denser OSG films, it is expected that insights in mechanical properties that are discussed in the following sections of this chapter should be directly applicable to ULK materials.

12.2 Channel Crack Failures

12.2.1 Film Thickness Effects

One of the most common forms of mechanical failure for brittle thin films under tensile stress is channel cracking. Figure 12.2 shows a cross-section scanning electron micrograph of a typical channel crack that developed in 3-μm thick OSG film that was deposited on a silicon wafer with residual tensile stress of 60 MPa. The image shows a through film thickness crack that propagated along the in-plane direction, i.e., across the face of the silicon wafer. The driving force of channel crack fracture is defined as the energy release rate (G) and expressed as shown in Equation 12.1 where σ, h, and $\bar{E}$ are the residual film stress, thickness, and the plane-strain elastic modulus of the thin film, respectively.

$$G = Z\frac{\pi\sigma^2 h}{2\bar{E}} \tag{12.1}$$

The parameter Z is a constant depending on the channel crack geometric and elastic constraint imposed by materials near the crack tip. Critical failure occurs when the energy release rate is greater than the fracture toughness (Γ) of the material. At strain energies less than the inherent fracture toughness of the material, it is still possible to observe crack growth depending on the material and crack tip reaction conditions. This type of fracture, commonly referred as subcritical crack growth or stress corrosion cracking, depends on the reactivity of the surrounding environment and typically occurs at slower rates than crack velocities measured under critical fracture conditions. Channel crack propagation velocity (V) can be expressed as shown in Equation 12.2, where k is the Boltzmann's constant, T is the absolute

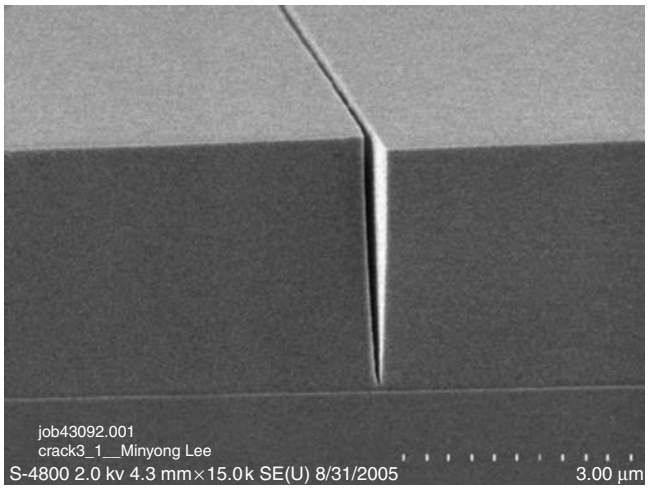


FIGURE 12.2 A scanning electron micrograph cross-section of an OSG channel crack.

temperature, N is the bond density, and V_0 is a reference velocity [6–8].

$$V = V_0 e^{(G/2NkT)} \tag{12.2}$$

Combining Equation 12.1 and Equation 12.2 results in the following equation

$$V = V_0 e^{(Z\pi\sigma^2 h/4\bar{E}NkT)} \tag{12.3}$$

Equation 12.3 shows that the crack growth rate depends exponentially on the film thickness (h) and the crack geometric/elastic constraint factor (Z) [9–11]. This equation also suggests that the crack propagation rate can be reduced by (1) increasing the film elastic modulus and/or (2) reducing the film stress. Cook and Liniger [6] demonstrated that the channel crack propagation rates of hydrogen silsequioxane (HSQ) and methyl silsequioxane (MSQ) thin films deposited on silicon indeed follow the velocity to film thickness relationships of Equation 12.3. Tsui et al. reported a similar dependence of crack growth velocity on thickness for OSG thin films [12,13].

The crack propagation rates of $k \sim 2.85$ Pe-CVD OSG low-k films in typical ambient conditions, 22°C and 45% relative humidity (RH), were measured as a function of film thickness and are shown in Figure 12.3 [12,13]. These films were characterized by plane-strain elastic modulus ($\bar{E}$) and film residual tensile stress (σ) of 8 GPa and 60 MPa, respectively. The data reported in Figure 12.3 show that the crack growth rate can be described by an exponential relationship with the film thickness (h) or film energy release rate (G) when it is slower than 100 µm/s. This is the characteristic fracture behavior in the reaction-controlled regime and can be expressed by Equation 12.1, as discussed above. In this region, the rate at which cracks propagate is controlled by chemical reaction rates at the crack tip and not the rate at which reactants can diffuse to the crack tip. With even thicker films, the exponential relationship between film thickness and crack velocity appears to no longer suffice, suggesting that this may corresponds to the onset of diffusion controlled channel cracking.

12.2.2 Extracting Materials Properties

To demonstrate the effects of low-k elastic modulus on the channel cracking behaviors, Tsui et al. [12,13] simplified Equation 12.3 to

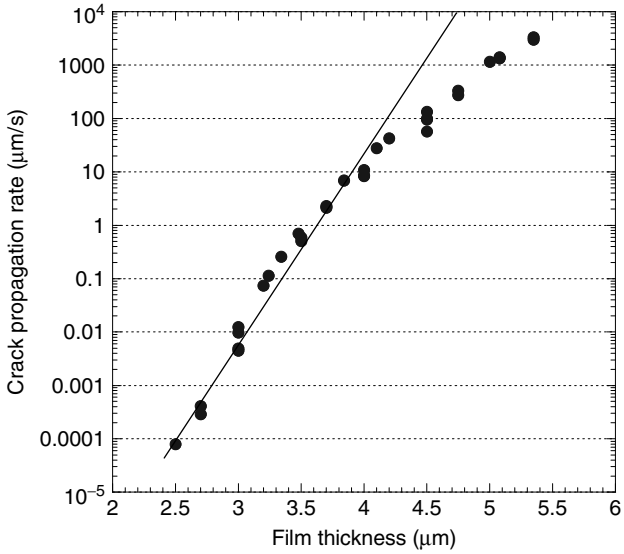


FIGURE 12.3 Crack propagation rates of organosilicate glass films in ambient air conditions as a function of film thickness.

$$\ln(V) = b_1 + b_2\left(h/\bar{E}\right) \tag{12.4}$$

where

$$b_1 = \ln(V_0) \tag{12.5}$$

and

$$b_2 = \frac{\pi Z \sigma^2}{4NkT}. \tag{12.6}$$

Assuming equivalent residual film stress (σ) and bond density (N) for films of varying thickness, but equivalent origin, b_1 and b_2 were obtained by curve fitting crack growth velocity data that were acquired as a function of film thickness. The data reported in Figure 12.3 corresponded to crack velocity measurements for OSG films of varying thickness, as measured in typical ambient conditions. Curve fitting this data yielded b_1 and b_2 materials constants reported in Table 12.2. As will be discussed later, the environment in which channel cracks grow has a significant effect on crack velocity in tensile films. Table 12.2 also includes b_1 and b_2 values extracted from similar crack propagation rate studies for OSG films soaked in deionized (DI) water [14]. Using Beuth's model [9], an elastic constraint factor, Z, of $\sim$0.743 is assumed for OSG films deposited on silicon. Equation 12.6 was used to calculate an OSG bond density (N), given that b_2 had been empirically fit from the aforementioned crack velocity vs. film

TABLE 12.2 Channel Crack Fracture Constants (b_1 and b_2) and Bond Density for OSG Films Deposited on Bare Silicon Substrates

	b_1 (m/s)	b_2 (N/m^3)	N (bonds/m^2)
Ambient air	−41.38	6.07×10^{16}	8.36×10^{18}
Deionized water	−33.57	5.91×10^{16}	8.58×10^{18}

thickness study. Tsui et al. [13] used the b_2 values measured in ambient and DI water to determine OSG bond density (N) of 8.4×10^{18} and 8.6×10^{18} bonds/m^2, respectively. Vlassak et al. [8] reported a similar bond density of 8.6×10^{18} bonds/m^2 using complementary subcritical four-point bend fracture techniques in ambient and a density of 8.4×10^{18} bonds/m^2 as measured by Rutherford backscattering spectroscopy. In summary, crack velocity experiments allowed one to define materials constants, b_1 and b_2, describing channel cracking for a set of OSG films, and thereby calculate bond density at the crack tip, a metric that was independently corroborated.

As previously mentioned, at the 90- and 65-nm technology nodes, the semiconductor industry widely adopted Pe-CVD OSGs as low-dielectric constant materials for high performance interconnects. A number of organosilane precursors have been developed to deposit $Si_xO_yH_z(CH_3)_u$ films with the varying stoichiometry (x:y:z:u) depending on the reaction conditions. It would be extremely useful to understand if extensive characterization of the one class of OSG films reported by Tsui et al. [13] that yielded materials properties b_1 and b_2 may be generalized to OSG films derived from different feed stocks and reaction conditions.

Using Equation 12.4 and the fracture constants $b_1 = -41/38$ m/s and $b_2 = 6.07\times10^{16}$ N/m^3, it was possible to construct a map of theoretical crack growth velocity as a function of film thickness for films of varying moduli (Figure 12.4). The dash and solid lines in the figure correspond to odd and even moduli, respectively. If the behavior report by Tsui et al. [13] is general, then one would expect that the crack velocity measurements for a set of different OSG films could be accurately superimposed on the map. In Table 12.3, the mechanical properties, film composition, and density of four Pe-CVD OSG films are reported as well as similar data for HSQ and MSQ, the two spin-on low-k films that were studied by Cook and Liniger [6]. As shown in Figure 12.4, experimental crack velocity data for these six films correlated well with the predicted crack velocity behavior, based on their plane-strain elastic modulus. This analysis indicated that the fracture constants b_1 and b_2 can accurately describe the channel cracking properties of four Pe-CVD OSG films. Furthermore, the applicability of this model to HSQ and MSQ thin-film

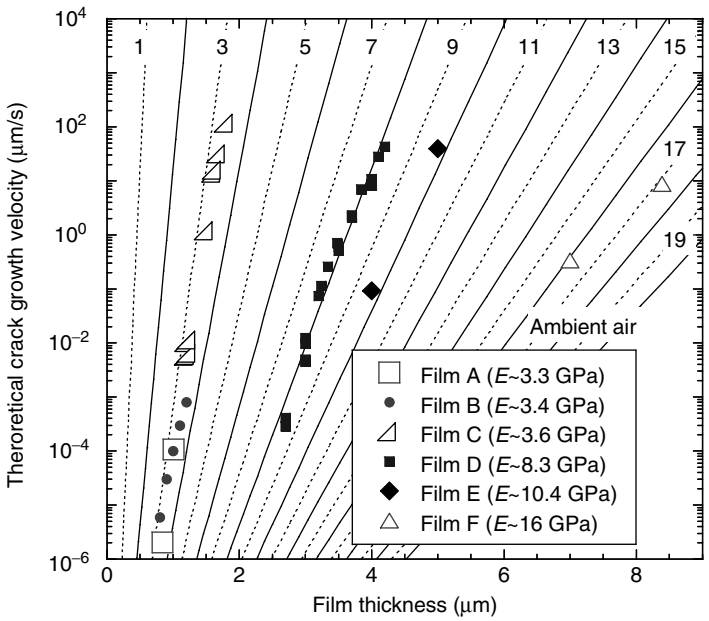


FIGURE 12.4 Theoretical and experimental crack propagation rate of low-k materials in ambient conditions as a function of film thickness.

TABLE 12.3 Mechanical Properties, Chemical Composition, and Density of Several Low-k Thin Films

	Mechanical		% at Composition[a]			
	E (GPa)	σ (MPa)	Si	C	O	Density[b] (g/cm^3)
Film A—hydrogen silsequioxane	3.3	60	40	0	60	Unknown
Film B—methyl silsequioxane	3.4	65	28.57	28.57	42.86	Unknown
Film C	3.5–3.7	59	37.24	15.71	47.06	1.27
Film D	8.0–8.5	60	37.6	14.6	47.8	1.43
Film E	8.9–11.8	60	37.3	21.93	40.77	1.4
Film F	15–17	55	39.19	19.74	41.02	1.44

[a] XPS
[b] X-ray reflectivity

channel cracking suggested that these materials constants may be applicable to a wider range of silicate-based low-k thin films.

The same methodology was used to characterize OSG films cracking in DI water, $b_1 = -33.57$ m/s and $b_2 = 5.91 \times 10^{16}$ N/m^3, and generate a theoretical crack growth velocity vs. thickness map (Figure 12.5). As before, dash and solid lines in the figure correspond to odd and even moduli, respectively. In addition to the materials listed in Table 12.2, crack velocity measurements reported by Grill et al. [15] are also included in this figure. As with the data collected under ambient conditions, there is a good correlation between predicted and experimental crack velocities for this wide range of OSG films tested in DI water, indicating that the DI water b_1 and b_2 values reported in Table 12.2 may also apply to a wide range of silicate-based low-k materials.

One of the important observations in both Figure 12.4 and Figure 12.5 is that at any given film thickness, the changes in crack growth rate are not linearly proportional to film modulus. This is best

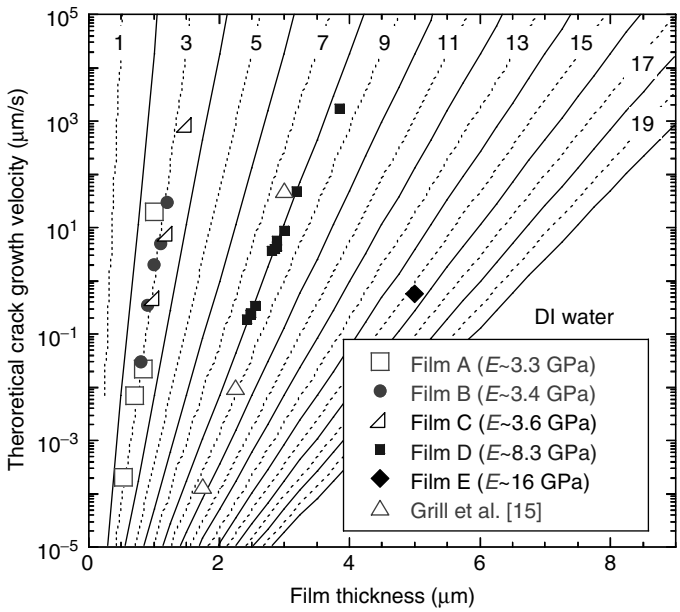


FIGURE 12.5 Theoretical and experimental crack propagation rate of low-k materials in deionized (DI) water as a function of film thickness.

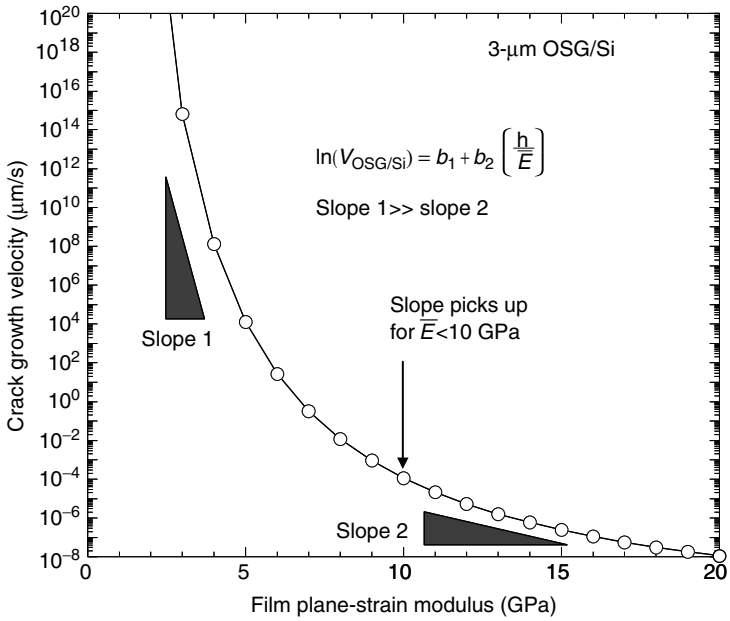


FIGURE 12.6 Theoretical ambient crack growth velocity of a hypothetical OSG film 3-μm thick with residual tensile stress of 60 MPa.

illustrated in Figure 12.6, where the calculated ambient crack growth rate of an arbitrary 3-μm thick low-k film with the characteristic residual stress of 60 MPa is plotted as a function of film modulus. For the sake of discussion, we assumed that the crack growth rate is within the reaction-controlled regime and used b_1 and b_2 values reported in Table 12.2. This model suggests that the cracks propagate very slowly for high modulus OSG films and the crack rate is insensitive to moduli variations. For instance, a material with modulus of 20 GPa would exhibit crack growth rates of approximately 1.0×10^{-8} μm/s. For more compliant materials, the crack propagation rate and its sensitivity to film thickness increase rapidly. A material with elastic modulus of 5 GPa is predicted to exhibit a crack propagation rate of 1.3×10^{4} μm/s. This represents a 10^{12} increase in crack velocity, in comparison with the stiffer film with a modulus of 20 GPa. Given the exponential relationship between plane-strain modulus and crack velocity, variations in modulus are predicted to have a significant effect on more compliant films. Specifically, a 1 GPa reduction in film modulus for the 5 GPa film will lead to a $\sim 10^{5}$ increase in crack propagation rate, assuming that the film fracture remains in the reaction-controlled regime. A similar 5% or 1 GPa elastic modulus reduction for a material with a modulus of 20 GPa will increase the crack propagation rate by only $\sim$150%.

Channel cracking, as observed with OSG thin films, is a potential failure mode that was not a concern in older technology nodes in which compressive SiO_2 or FSG was used as intermetal dielectrics (IMD). The proceeding discussion highlights aspects of the channel cracking problem and provides a framework for predicting the cracking behavior of low-k OSG materials in ambient and DI water. For low elastic moduli films, it is also clear that changes in stiffness, possibly due to processing variation, could affect significant changes in crack growth behavior. Looking forward to future technology nodes, ULK OSG materials are likely to be even more prone to this type of mechanical failure due to their low elastic moduli and reduced bond density. This could be investigated with similar experiments using a prototypical ULK material, yielding new b_1 and b_2 values that are characteristic of the low-density ULK materials.

12.3 Elastic Constraint Effects

The driving force for channel crack fracture is characterized by the energy release rate (G) model described by Equation 12.1. This equation shows that the crack propagation rate has an exponential relationship with material properties and geometric variables. In the previous sections, we discussed the effects of the film thickness (h) and the film plane-strain elastic modulus ($\bar{E}$) on the channel cracking behavior of low-k OSG films deposited on a stiff silicon substrate. In this section, we will expand our investigation of channel cracking to examine the influence of elastic properties of the substrate on channel cracking and by extension, channel cracking behavior of OSG films deposited in multilayer film stacks.

Equation 12.1 shows that the driving force for fracture in thin films is linearly proportional to the elastic constraint factor (Z). Beuth [9] and Vlassak [11] demonstrated that this value is based on the plane-strain Dundurs parameters [16], where ν represents the Poisson's ratio and μ the shear modulus (Equation 12.7). In Equation 12.7, the subscripts represent substrate (s) or film (f) properties.

$$\alpha = \frac{\bar{E}_f - \bar{E}_s}{\bar{E}_f + \bar{E}_s}, \qquad \beta = \frac{\mu_f(1-2\nu_s) - \mu_s(1-2\nu_f)}{2\mu_f(1-\nu_s) + 2\mu_s(1-\nu_f)} \tag{12.7}$$

Values for Z corresponding to different elastic modulus mismatches between the film and the substrate were reported by Beuth [9] and are plotted as a function of α with $\beta = \alpha/4$ for a film on an elastic half space in Figure 12.7. The figure reveals that Z value does not change significantly when α is negative, i.e., the substrate is stiffer than the film. In contrast, if the film is stiffer than the substrate, the Z value increases rapidly with α.

The significance of this strong film/substrate elastic mismatch sensitivity can be thought of as follows. A compliant OSG film of a given thickness and residual stress deposited on a stiff substrate may not fracture, because the energy release rate is below the threshold of crack formation. However, when the

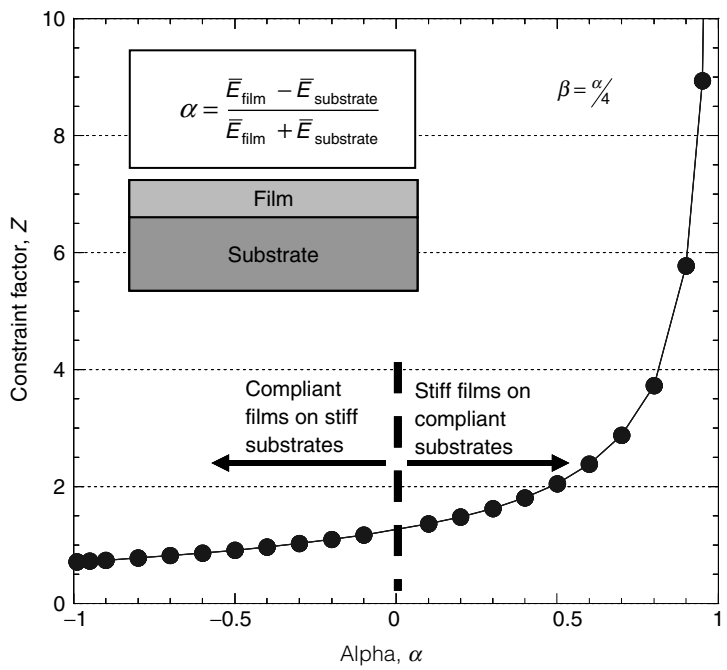


FIGURE 12.7 Plot of the elastic constraint factor (Z) as a function of Dundurs parameter (α).

TABLE 12.4 Mechanical Properties of the Buffer-Layer Materials

	Elastic Modulus (GPa)	Plane-Strain Modulus (GPa)	Residual Stress (MPa)
Si[a]	163	172	NA
SiN_x	155	165	-125 ± 5
SiO_2	70	75	-135 ± 5
OSG	8	8.5	60 ± 2
Organic polymer	3.5	3.7	60 ± 2
Low-density carbon-doped silicon oxide	3.5	3.7	60 ± 3

[a] Polycrystalline values calculated from single-crystal elastic constants.

identical film, i.e., the same physical properties, is bonded to a compliant substrate, mechanical failures may occur readily because of the increase in constraint factor (Z) and by extension, the energy release rate.

In one sense, the proceeding discussion can be simplified as the semiconductor industry typically focuses on using stiff silicon substrates for fabricating all high performance devices, i.e., α is negative. Upon closer examination, however, the situation is more complicated as low-k materials are integrated with other dielectric materials, i.e., Pe-CVD silicon nitride or silicon carbide, and metals and in all cases the elastic properties of these materials differ from those of OSGs. In this section, we will expand our examination of channel cracking to comprehend the impact of elastic properties of the underlayer or buffer layers in a multilayered system on the constraint factor and the film energy release rate.

Tsui et al. [17,18] studied the channel cracking behaviors of Pe-CVD OSG films deposited on five different dielectric materials with various elastic properties: silicon nitride (SiN_x), SiO_2 deposited from tetra-ethoxysilane, low-density carbon-doped silicon oxide (LD-CDO) deposited from octamethyl cyclotetrasiloxane, and a poly-aromatic polymer. The nanoindentation elastic moduli of these films are listed in Table 12.4. Silicon oxide and silicon nitride films are more compliant with the smaller elastic moduli than the silicon substrate, but greater than the brittle OSG film. The other two buffer-layer materials (LD-CDO and polymer films) are approximately two times more compliant than OSG.

The channel crack growth velocity of 3.25-μm thick OSG films deposited on silicon nitride or SiO_2 buffer layers are plotted as a function of their normalized thickness in Figure 12.8. The crack growth rate of a 3.25-μm thick OSG film deposited on bare silicon is included for reference. The data reported in Figure 12.8 showed that cracks grow faster when OSG is deposited on these buffer layers than when deposited directly on silicon. These observations can be understood in light of a reduction in elastic constraint due to presence of interposing buffer layers with smaller elastic moduli, relative to silicon. According to Beuth's calculations [9], the buffer layer will increase the effective values of Dundurs parameter (α), constraint factor (Z), and the OSG film energy release rate (G)—the driving force for channel crack propagation. Since the SiO_2 film is a more compliant material than SiN_x or the silicon substrate, the amount of constraint imposed on the OSG film is reduced and produces greater driving force for fracture; i.e., faster crack growth velocity. Data reported in Figure 12.8 also show that the crack velocity increases with the buffer-layer thickness. As the thickness of the buffer layer increases the majority of the channel crack elastic field is contained within the buffer layer, affecting reduced elastic constraint from the substrate and an increased energy release rate. With sufficiently thick buffer layers, the crack driving force (energy release rate) approaches that limiting case in which the OSG film was deposited on a substrate with the same elastic properties as the buffer layer. The maximum theoretical velocity limits for OSG films deposited on SiO_2 and SiN_x buffer layers are shown in Figure 12.8.

Based on the work of Beuth [9], one would expect a significant increase in OSG crack velocity, if the buffer layer interposed between the low-k dielectric film and the silicon substrate was characterized by modulus that was less than that of OSG. In the study by Tsui and co-workers [12,13], thin films of LD-CDO or poly-aromatic polymer films served this purpose as both of these materials have elastic moduli that are smaller than the silicon substrate and OSG. Crack propagation rate data for 3-μm thick

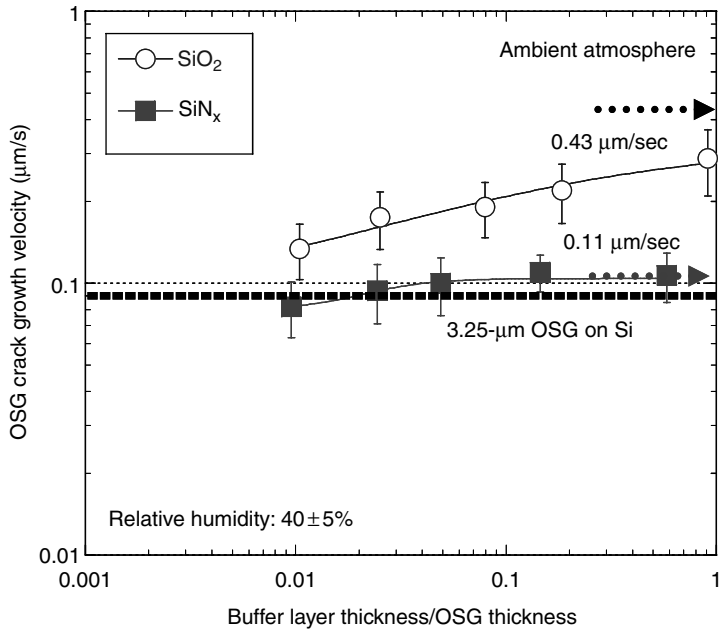


FIGURE 12.8 Plot of the OSG crack propagation rate as a function of normalized buffer-layer thickness.

OSG films deposited on these two buffer layers and on bare silicon substrate is shown in Figure 12.9. Crack velocities for the bilayered systems were significantly faster than similar OSG films deposited directly on a silicon substrate. In fact, the crack velocity of OSG on 150-nm thick LD-CDO buffer layer (5% of the OSG thickness) is 10 times of the crack speed on silicon substrate. With a 500-nm LD-CDO underlayer or 16% of the OSG thickness, the crack rate is more than 10^5 times greater than on a bare silicon substrate. As predicted by theory, the crack growth rate is more sensitive to buffer-layer thickness when the buffer layer is more compliant (α is positive) than the OSG film.

It is instructive to compare the constraint effects on channel cracking across technology nodes as this can provide insight into cracking failure susceptibility of interconnects and provide guidelines for future material selection [12,13]. Channel crack growth rates of 3-μm thick OSG film deposited on multilayered BEOL structures made from different IMD materials are shown in Figure 12.10. Each sample consists of a different number of alternating layers of 300-nm IMD and 60-nm etch-stop (ES) layers. The ES material is silicon carbonitride with dielectric constant of 5.0 and elastic modulus of 100 GPa. SiO_2, OSG, and LD-CDO were chosen as representative IMDs for the 180-, 90-, and 45-nm technology nodes, respectively. Figure 12.10 shows that the crack growth rate increased with an increasing number of bilayers beneath the OSG films. This is expected because all of the buffer-layer materials included in this study have elastic moduli that are smaller than the silicon substrate. The greater the number of alternating IMD/ES buffer layers, the more the elastic constraint from the silicon substrate was reduced and the constraint factor values (Z), film energy release rate (G), and the crack propagation rate increased. This effect is dramatic for low moduli materials such as LD-CDO as this study showed a 10^5 increase in OSG crack velocity, compared with SiO_2, for a composite of 10 pairs of IMD/ES film stacks.

These results demonstrate the potential for mechanical failure risks due to channel cracking for future interconnect technology nodes where advanced low-*k* materials will have significantly lower elastic modulus than the dense silicon oxide or the OSG class of low-*k* materials that have been successfully integrated at the 90-nm technology node.

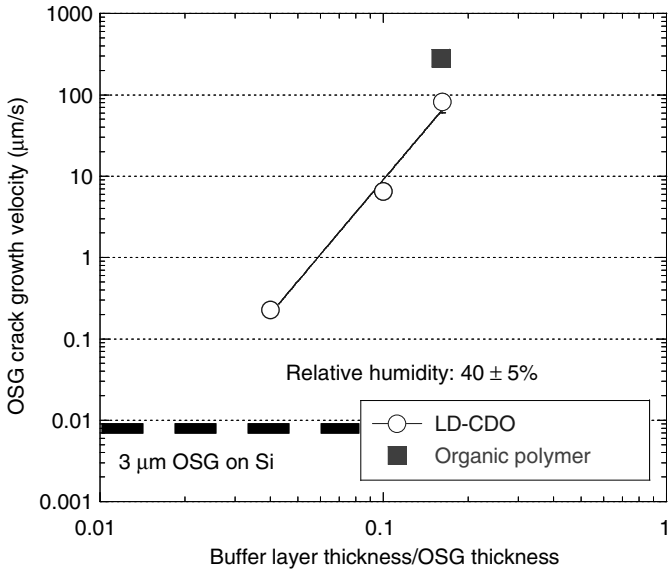


FIGURE 12.9 Plot of the OSG crack propagation rate as a function of normalized buffer-layer thickness.

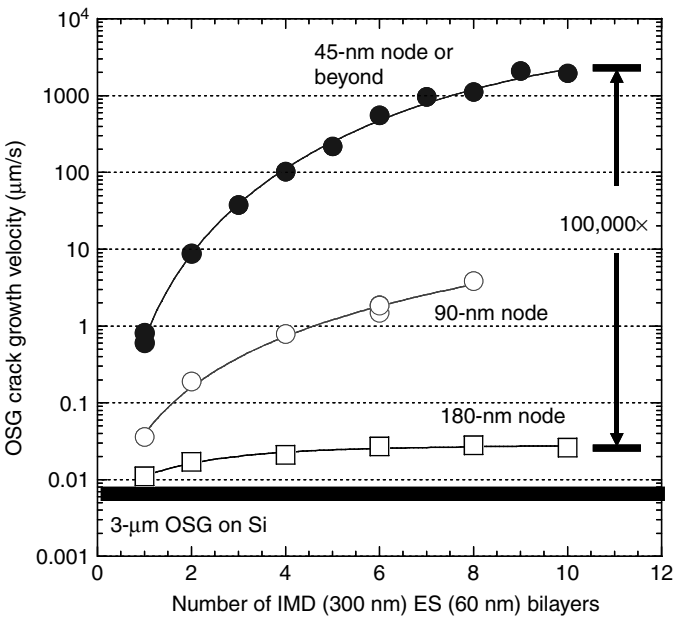


FIGURE 12.10 Organosilicate glass crack growth velocity for different numbers of IMD/ES bilayers.

12.4 Pattern Layout Effects

In addition to elastic constraint, environment, and reactant absorption effects on low-*k* film fracture, the propensity for dielectric cracking can also be influenced by the pattern layout, the arrangement of metal lines, and the dielectric spaces in the interconnect. An additional crack driving force can originate from the thermal expansion coefficient mismatch between the copper lines and the surrounding low-*k* dielectrics. Liu et al. [19] and Ambrico et al. [20] used the finite element analysis (FEA) and analytical methods to study the effects of copper line geometry, such as height, width, and line spacing, on the low-*k* dielectric elastic-stain-energy near metal structures.

Liu et al. [19] studied the energy release rate for channel cracks propagating under two different scenarios. The first model was for cracks propagating in between two interconnect lines, while the second case was for cracking in a low-*k* interlayer dielectric layer directly above the metal lines. The schematic drawings of these two FEA models are shown in Figure 12.11 and Figure 12.12. Channel cracks were assumed to propagate parallel to the metal lines, where parameters h_c and w correspond to the metal line height and the spacing between interconnect lines, respectively. Liu et al. [19] characterized the crack driving force using film energy release rate (G).

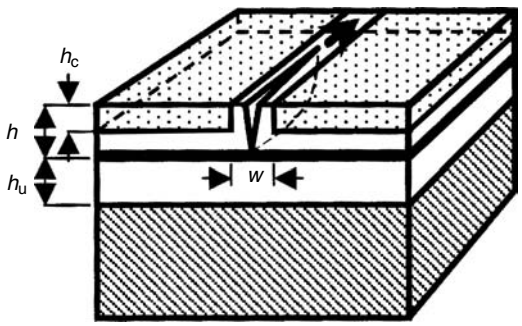


FIGURE 12.11 Schematic drawing of channel crack modeled by Liu et al. (Liu, X. H. et al., *Advanced Metallization Conference 2004 (AMC 2004)*, ed. Erb, D., Ramm, P., Masu, K., and Osaki, A., Materials Research Society, Warrendale, PA, 2005, 361.)

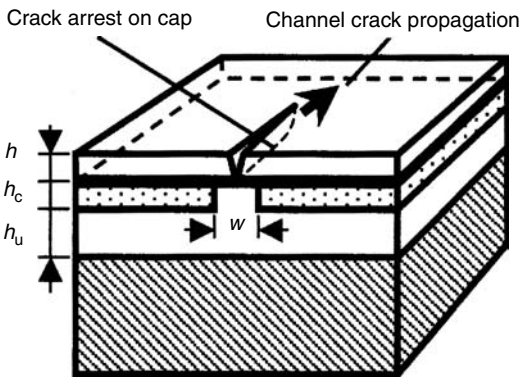


FIGURE 12.12 Schematic drawing of channel crack modeled by Liu et al. (Liu, X. H. et al., *Advanced Metallization Conference 2004 (AMC 2004)*, ed. Erb, D., Ramm, P., Masu, K., and Osaki, A., Materials Research Society, Warrendale, PA, 2005, 361.)

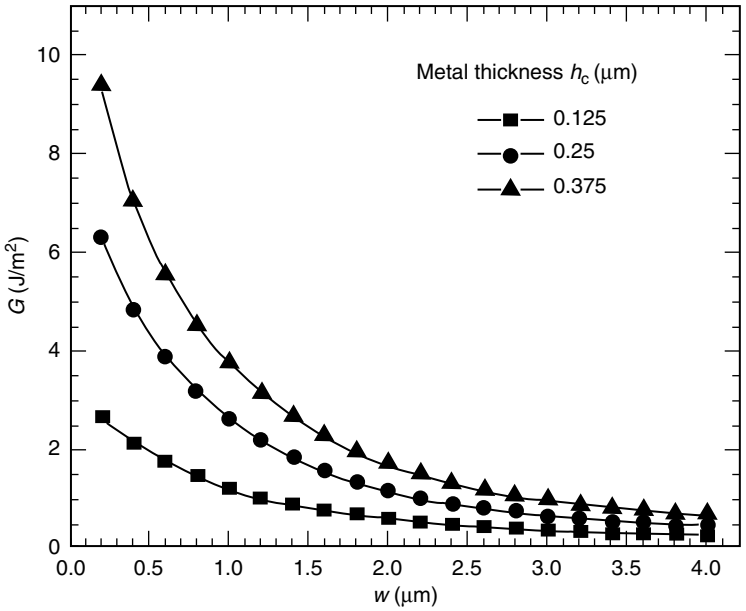


FIGURE 12.13 Film strain-energy release rate of as a function of metal spacing (w) and the metal thickness (h_c).

Results for the first FEA model, channel cracking between two metal lines, are plotted as a function of line spacing (w) for three different metal thicknesses (h_c) in Figure 12.13. The data reported in this figure showed that the driving force for channel cracking was large when the distance between metal lines was small or the metal interconnect lines were thick. The film energy release rate reduced rapidly and approached the blanket film value as the spacing between the lines increased. The propensity for dielectric fracture in the first BEOL geometry was also decreased by reducing the metal thickness.

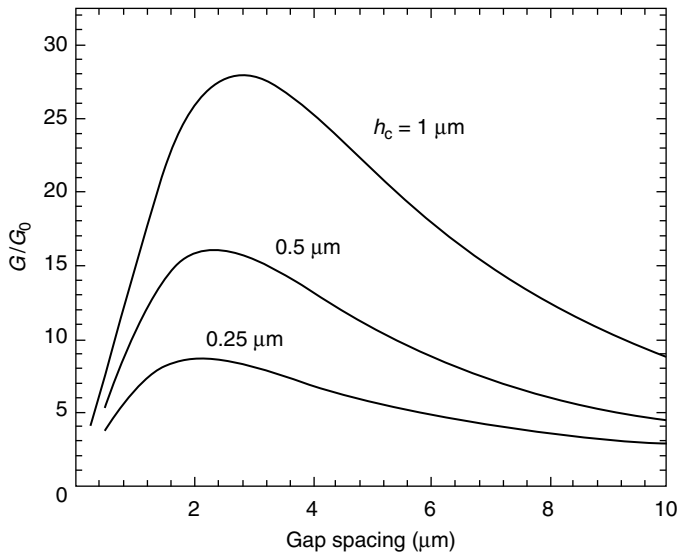


FIGURE 12.14 Film strain-energy release rate of as a function of metal gap spacing (w) and the metal thickness (h_c).

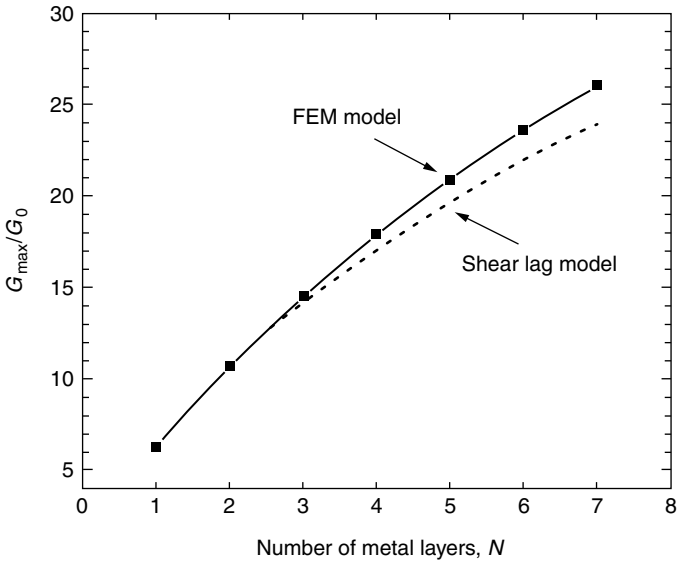


FIGURE 12.15 Film strain-energy release rate of as a function of metal layers.

The FEA low-*k* film energy release rate results from the second FEA structure, channel cracking in a film deposited on top of metal lines, are shown in Figure 12.14. Liu et al. reported that the crack driving force was small, when the metal lines were close to each other, i.e., a small gap spacing. The dielectric energy release rates increased with the metal spacing and reached a maximum value before reducing back to the blanket film value. This study demonstrated that there is a critical metal spacing value for each metal thickness where the driving force for channel crack formation is the highest and should be avoided. The effects of elastic constraint on the low-*k* fracture were also simulated in Liu's work. The normalized energy release rate as a function of metal layers (N) is plotted in Figure 12.15. The data reported in this figure showed that the crack driving force increased rapidly with the number of metal layers and could be described by a modified shear lag model developed by Suo [21]. The film energy release rate at metal seven is five times greater than the value at the first metal layer.

The results presented in this section demonstrate the importance of layout design variables, including metal thickness and spacing that could have a significant impact on channel cracking in materials that are susceptible to this type of mechanical failure.

12.5 Environmental Effects

12.5.1 Interfacial Adhesion

The effects of environment on bulk silicate glass fracture were studied by Wiederhorn et al. using double countilever beam (DCB) techniques. They showed that the crack propagation rate was affected by the partial pressure of water [22,23], solution pH [24], and the atmospheric pressure in the environment [25]. Furthermore, Wiederhorn also reported that the silica bond fracture energy was related to the water partial pressure (p_{H_2O}) in the environment as shown in Equation 12.8, where parameters A and B are constants.

$$\Gamma = B - A \ln(p_{H_2O}) \tag{12.8}$$

Recently, Guyer and Dauskardt [26], Lane et al. [27,28], Lin et al. [29,30], and Vlassak et al. [8] used DCB and four-point bend techniques to evaluate environmental effects on the fracture properties of

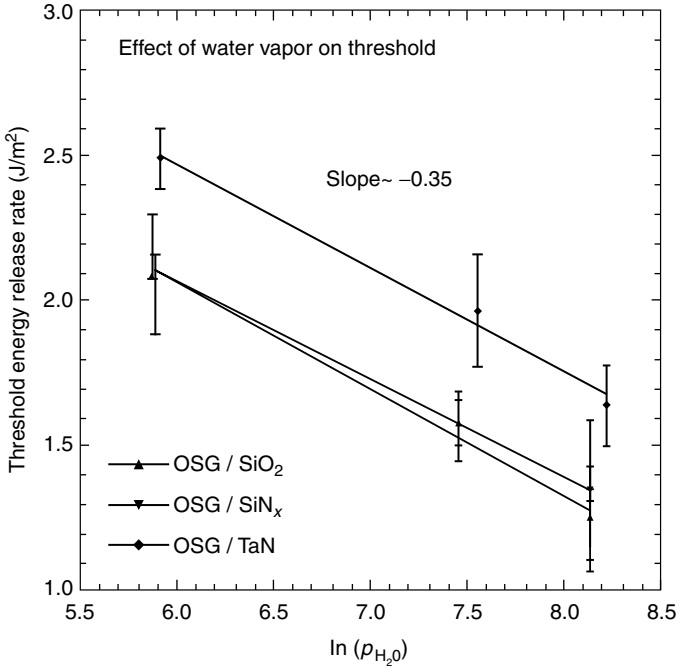


FIGURE 12.16 Threshold fracture energy release rate for OSG with three different capping materials in different humidity environments.

low-k materials. Their results agreed with Wiederhorn's work showing that the crack propagation rates increased with the logarithmic partial pressure of water and followed the empirical relationship described by Equation 12.8.

Figure 12.16 shows the threshold adhesion strength of Pe-CVD OSG films capped with three different materials: silicon oxide, silicon nitride, and tantalum nitride (TaN). These measurements were made using four-point bend testing operating under different humidity conditions [8]. The figure shows that adhesion, as measured by the threshold energy release rate, decreases with increasing water concentration in the environment and can be described by Equation 12.8. Curve fitting of the results show the slope (A) of all three samples to be similar, while the TaN-capped sample had a significantly different y-intercept (B) than the SiO_2- and SiN_x- capped films. Lin et al. [29,30], Guyer and Dauskardt [26], and Vlassak et al. [8] also reported that the fracture threshold and the subcritical crack rates increased with solution pH. This is best illustrated in Figure 12.17 where the OSG/TaN interface crack growth velocity is plotted as a function of the energy release rate for tests performed in media of differing pH (pH 3, 7, and 12). This study showed that the fracture threshold decreased in more alkaline environments and the crack propagation rate increased with pH when cracking occurred in the reaction-controlled regime.

12.5.2 Channel Cracking

Moisture effects on channel crack propagation behavior of spin-on HSQ and MSQ thin films were studied by Cook et al. [6]. They investigated cracking velocity as a function of chemical composition and thermal cure processes. Their results showed that cracks propagated faster in the moist environments than in the dry atmospheric conditions, as predicted by Wiederhorn [22]. Recently, Jacques et al. [31] measured channel crack propagation rates in different environments for Pe-CVD OSG thin films. They studied the effects of water partial pressure, organic solvents, and pH on crack velocity in the reaction-controlled regime. In Figure 12.18, crack rate data are reported as a function of energy release rate for

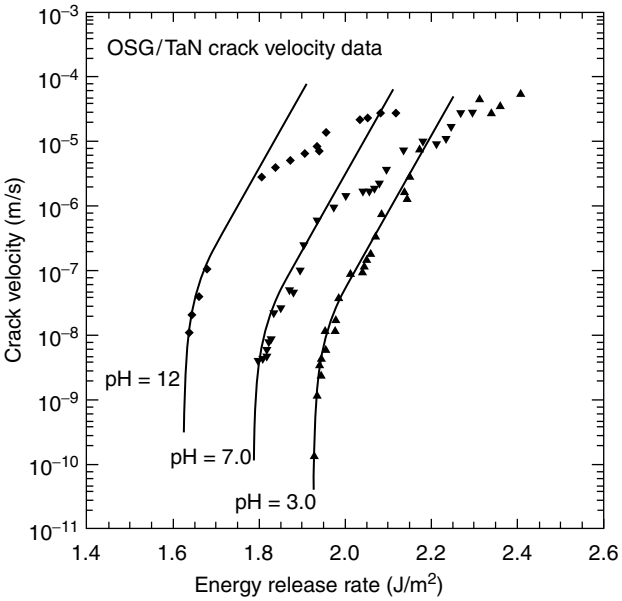


FIGURE 12.17 Subcritical crack propagation rate of tantalum nitride/OSG interfaces in different pH environments.

3-μm thick OSG films exposed to different environments including: purging with N_2 (23°C and 10% RH), ambient (23°C and 45% RH), soaking in DI water. As predicted by Equation 12.8, an increase in water partial pressure increased the crack propagation velocity, suggesting that increased water concentration reduced the cohesive strength (Γ) of silicon oxide bonds at the crack tip. At an arbitrary energy release rate of 1.4 J/m^2, this study demonstrated that 3-μm thick OSG films cracked $\sim 10^6$ faster in water than when stored in a nitrogen-purged environment with less than 10% RH. When studied under ambient conditions, the crack rate was approximately 1000 times greater than the rate measured when

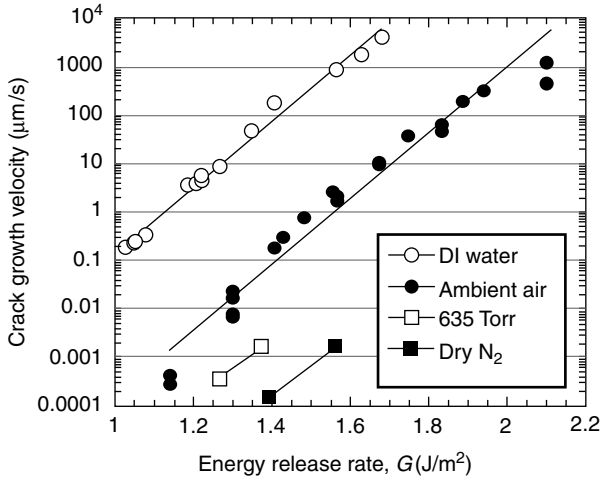


FIGURE 12.18 Channel crack propagation rate of OSG films in different environments.

samples were stored in a nitrogen-purged environment. In the same study, Jacques et al. [31] measured crack velocity of similar OSG films stored at subatmospheric pressures. Specifically, they reported that the crack velocity of 3-μm thick OSG films decreased from 0.01 to 0.0004 μm/s when samples were stored in a chamber at a reduced pressure of 625 Torr. At an ultra-low pressure of 1×10^{-9} Torr, the crack growth rate decreased further to less than 10^{-6} μm/s or 0.01 A/s.

The information presented in this and previous sections highlights the impact of environment on crack propagation in low-*k* OSG films. The sensitivity of OSG materials to moisture induced reductions in threshold adhesion strength or channel cracking has potential implications for fabricating interconnects with these materials and their subsequent packaging. Furthermore, the mechanical failure modes discussed, interfacial delamination and channel cracking, are likely to be an even greater concern for porous OSG thin films that are being considered as candidate ULKs for the 32-nm technology node and beyond.

12.5.3 Diffusion Studies

The OSG materials that are being integrated at the 90- and 65-nm technology nodes are characterized by low density (Table 12.3) but are not truly porous. They can be thought of as open network glasses that include voids/pores on the nanometer scale. As such, water can diffuse into these glasses and affect mechanical failures including interfacial delamination and channel cracking. Diffusion coefficients for silica-based and aromatic polymer low-*k* materials are listed in Table 12.5 [32–35]. Moisture diffusivity in bulk dense silicate glass and quartz are also included for comparison. As expected, diffusion coefficients for low-*k* materials are at least 10^9 times greater than those of dense silicate glass. Lin et al. [29,30], Guyer and Dauskardt [26], and Vlassak et al. [8] reported that water concentration at the crack tip environment can affect fracture properties of OSG thin films. Given that this class of materials readily absorb water and that water at a steady-state concentration will affect stress corrosion cracking, it is instructive to examine the effects of exposure time to water on interfacial adhesion.

Recently, Tsui et al. [36,37] and Lin et al. [39] performed a series of experiments to investigate temporal effects of reactant exposure on adhesion and fracture properties of low-*k* OSG films capped with tantalum (Ta), TaN, silicon nitride, and SiO_2 films. They characterized this effect by measuring the four-point bend adhesion strength of these interfaces after exposing samples to aqueous solutions for varying times. X-ray photoelectron spectroscopy (XPS) chemical analysis of the fractured surfaces showed failures within the OSG film for Ta-, TaN_x-, and SiN_x- capped samples, approximately 5–10 nm from the interface. For OSG/SiO_2 samples, XPS data showed that delamination occurred at the interface.

The results of the study are shown in Figure 12.19, where adhesion strength is plotted as a function of time that the sample was submersed in DI water at 24°C. Of these film stacks, the Ta/OSG interface was the strongest. Furthermore, the interfacial strength of all four samples degraded with water exposure time. After 1 week, adhesion strength reduced to less than 70% of what was originally measured. Also shown in Figure 12.19 is data for SiN- and SiO_2- capped OSG films that were baked at 120°C for 16 h showing that the interfacial strength was recovered, indicating that the degradation mechanism was related to the presence of water in the OSG films.

TABLE 12.5 Diffusivity of Moisture in Silica-Based and Aromatic Polymer (Silk) Dielectrics

	Diffusivity at Room Temperature (cm^2/s)	Temperature Range (°C)	Reference
Quartz	4.5×10^{-20} (Extrapolated)	125–200	32
Silica glass	4.5×10^{-17} (Extrapolated)	125–200	33
Plasma-enhanced chemical vapor deposition silica	$(7\pm2)\times10^{-17}$	8–90	35
SiLK™	1.7×10^{-6}	25	34

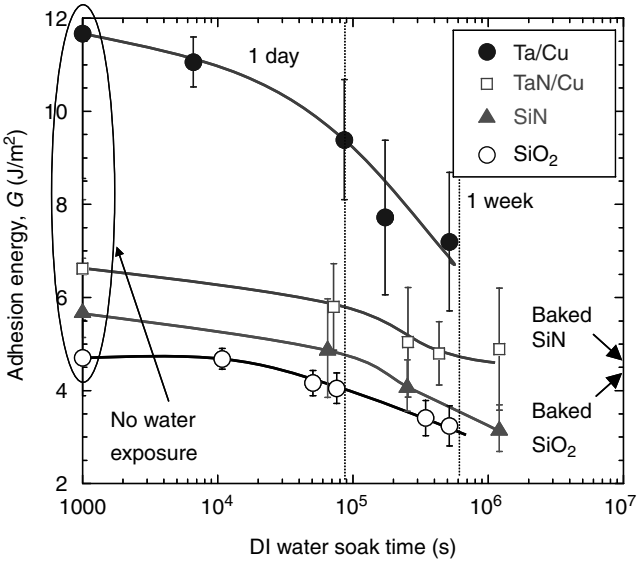


FIGURE 12.19 Plot of adhesion strength between OSG and different capping materials as a function of DI water exposure time.

Tsui and co-workers included in their study an evaluation of moisture exposure time on interfacial adhesion for SiO_2/OSG film stacks in which the OSG film thickness was varied from 0.42 to 3 μm [36,37]. As shown in Figure 12.20, all samples experienced a similar degradation in adhesion strength that was independent of film thickness. The as-prepared sample without water exposure was characterized by adhesion strength of ~5.5 J/m^2 but reduced to a plateau value near 2.7 J/m^2 after soaking in water for 1 month. The eventual stabilization of adhesion strength after the long exposure time is likely to correspond to the moisture saturation in OSG. Since the degradation mechanism is water

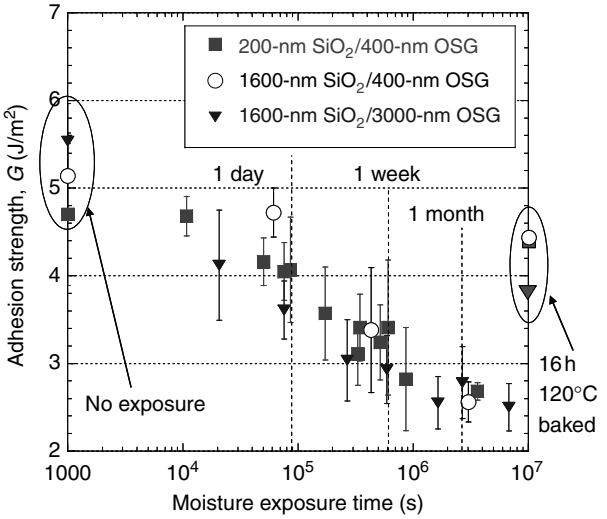


FIGURE 12.20 Adhesion strength of silicon dioxide and OSG interfaces as a function of the water exposure time.

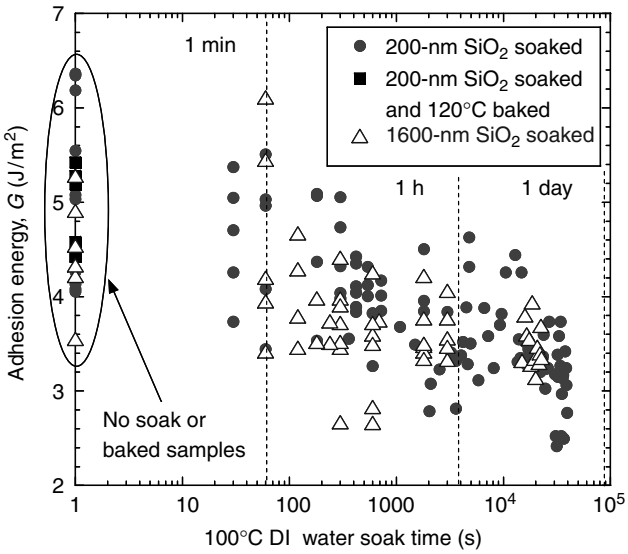


FIGURE 12.21 Adhesion strength of silicon dioxide and OSG interfaces as a function of 100°C water exposure time.

diffusion-driven, it is important to note that this saturation value and the rate of adhesion degradation can change with the water temperature and its partial pressure in the environment. Tsui et al. [37] examined this effect by submerging SiO_2/OSG samples in the 100°C DI water for different periods of time and performing critical four-point bend adhesion test at the ambient conditions afterward. Their results are shown in Figure 12.21 with the "dry" adhesion value of $\sim$5.2 J/m^2 and reduce rapidly with boiling water exposures to less than 3 J/m^2. The rate of adhesion degradation rate in boiling water is approximately 20 times faster than when soaked in water at room temperature.

From a semiconductor manufacturing perspective, this strong dependence of adhesion strength on moisture absorption could be significant. When an OSG film is capped, the dry adhesion value is above the threshold fracture value and no mechanical failure will occur. If the sample is exposed to a wet clean process, such as post-etch/ash clean, or simply exposed to an atmosphere which contains moisture, interfacial strength will degrade with time. If the adhesion strength is below the film energy release rate, mechanical failures like delamination begin to occur. For low-k OSG films, this has not been a significant issue. For ULK materials that are more prone to mechanical fracture, this should be viewed as a significant risk that needs to be comprehended in developing a high yielding process flow.

12.6 Conclusion

In this chapter we have detailed mechanical reliability issues associated with OSGs, a class of low-dielectric constant materials that have been widely and successfully integrated in high performance interconnects at the 90- and 65-nm technology nodes. We have reviewed factors affecting channel cracking, including film thickness effects, elastic constraint effects, and pattern layout effects. In addition, we have reviewed efforts to extract materials constants from crack propagation studies. The impact of environment on interfacial adhesion and channel cracking has also been reviewed. Looking forward to the 45-nm technology node, and beyond, it is likely that further reduction in bulk dielectric constant is likely to be achieved through the introduction of porosity into OSG-like films, yielding ULK films. The materials science issues that affected dense OSG materials (k=2.7–3.0) that were discussed in this chapter are likely to be a significant concern with these lower density, lower moduli ULK materials.

References

1. Ray, G. W. "Alternative Interlayer Dielectrics." In *Handbook of Semiconductor Manufacturing Technology*, edited by Y. Nishi, and B. Doering, New York: Marcel Dekker, 2000, chap. 12.
2. Lin, Y., T. Y. Tsui, and J. J. Vlassak. "Octamethylcyclotetrasiloxane-Based, Low-permittivity Organosilicate Coatings." *J. Electrochem. Soc.* 153 (2006): F144.
3. Wang, L., et al. "Nanoindentation Analysis of Mechanical Properties of Low to Ultralow Dielectric Constant SiCOH Films." *J. Mater. Res.* 20 (2005): 2080.
4. Grill, A., and V. Patel. "Ultralow-*k* Dielectrics Prepared by Plasma-Enhanced Chemical Vapor Deposition." *Appl. Phys. Lett.* 79 (2001): 803.
5. Grill, A., and D. A. Neumayer. "Structure of Low Dielectric Constant to Extreme Low Dielectric Constant SiCOH Films: Fourier Transform Infrared Spectroscopy Characterization." *J. Appl. Phys.* 94 (2003): 6697.
6. Cook, R. F., and E. G. Liniger. "Stress-Corrosion Cracking of Low-Dielectric-Constant Spin-on-Glass Thin Films." *J. Electrochem. Soc.* 146 (1999): 4439.
7. Toivola, Y., J. Thurn, and R. F. Cook. "Structural, Electrical, and Mechanical Properties Development during Curing of Low-*k* Hdrogen Silsesquioxane Films." *J. Electrochem. Soc.* 149 (2002): F9.
8. Vlassak, J. J., Y. Lin, and T. Y. Tsui. "Fracture of Organosilicate Glass Thin Films: Environmental Effects." *Mater. Sci. Eng. A* 391 (2005): 159.
9. Beuth, J. L. Jr. "Cracking of Thin Bonded Films in Residual Tension." *Int. J. Solids Struct.* 29 (1992): 1657.
10. Hutchinson, J. W., and Z. Suo. "Mixed-Mode Cracking in Layered Materials." *Adv. Appl. Mech.* 29 (1992): 63.
11. Vlassak, J. J. "Channel Cracking in Thin Films on Compliant Substrates." *Int. J. Fracture* 119 (2003): 299.
12. Tsui, T. Y., et al. Effects of Elastic Modulus on the Fracture Behavior of Low-Dielectric Constant Films. Vol. 8. San Fransisco: Presented at International Interconnect Technology Conference, June 6–8, 2005, 3.16.
13. Tsui, T. Y., A. J. Griffin, Jr., R. Fields, J. M. Jacques, A. J. McKerrow, and J. J. Vlassak. "The Effect of Elastic Modulus on Channel Crack Propagation in Organosilicate Glass Films. " *Thin Solid Films* 515(2006): 2257–61.
14. Jacques, J. M., et al. "Environmental Effects on Crack Characteristics for OSG Materials." In *Thin Films—Stresses and Mechanical Properties XI*, edited by T. E. Buchheit, A. M. Minor, R. Spolenak, and K. Takashima, *Materials Research Society Symposium Proceedings*. Vol. 875, O10.6, Warrendale, PA, 2005.
15. Grill, A., et al. *Optimization of SiCOH Dielectrics for Integration in a 90 nm CMOS Technology* San Francisco, CA: Presented at the 2004 International Interconnect Technology Conference, June 7–9, 2004, 54.
16. Dundurs, J. "Edge-Bonded Dissimilar Orthogonal Elastic Wedges." *J. Appl. Mech.* 36 (1969): 650.
17. Tsui, T. Y., A. J. McKerrow, and J. J. Vlassak. "Constraint Effects on Cohesive Failures in Low-*k* Dielectric Thin Films." In *Materials, Technology and Reliability of Advanced Interconnects—2005*, edited by P. R. Besser, A. J. McKerrow, F. Iacopi, C. P. Wong, and J. J. Vlassak, *Materials Research Society Symposium Proceedings*. Vol. 863, B 4.1, Warrendale, PA, 2005.
18. Tsui, T. Y., A. J. McKerrow, and J. J. Vlassak. "Constraint Effects on Thin Film Channel Cracking Behavior." *J. Mater. Res.* 20 (2005): 2266.
19. Liu, X. H., et al. "Low-*k* BEOL Mechanical Modeling." In *Advanced Metallization Conference 2004 (AMC 2004)*, edited by D. Erb, P. Ramm, K. Masu, and A. Osaki, 361. Warrendale, PA: Materials Research Society, 2005.
20. Ambrico, J. M., E. E. Jones, and M. R. Begley. "Cracking in Thin Multi-Layers with Finite-Width and Periodic Architectures." *Int. J. Solids Struct.* 39 (2002): 1443.
21. Suo, Z. "Reliability of Interconnect Structures." In *Comprehensive Structural Integrity*, edited by W. Gerberich, and W. Yang, Vol. 8, 265. Amsterdam: Elsevier, 2003.

22. Wiederhorn, S. M. "Influence of Water Vapor on Crack Propagation in Soda-Lime Glass." *J. Am. Ceram. Soc.* 50 (1967): 407.
23. Wiederhorn, S. M., et al. "Effects of Water and Other Dielectrics on Crack Growth." *J. Mater. Sci.* 17 (1982): 3460.
24. Wiederhorn, S. M., and H. Johnson. "Effect of Electrolyte pH on Crack Propagation in Glass." *J. Am. Ceram. Soc.* 56 (1973): 192.
25. Wiederhorn, S. M., et al. "Fracture of Glass in Vacuum." *J. Am. Ceram. Soc.* 57 (1974): 336.
26. Guyer, P. E., and R. H. Dauskardt. "Fracture of Nanoporous Thin-Film Glasses." *Nat. Mater.* 3 (2004): 53.
27. Lane, M., et al. "Subcritical Debonding of Multilayer Interconnect Structures: Temperature and Humidity Effects." In *Materials Reliability in Microelectronics IX*, edited by C. A. Volkert, A. H. Verbruggen, and D. Brown, *Materials Research Society Symposium Proceedings*. Vol. 563, 251, Warrendale, PA, 1999.
28. Lane, M. W., J. M. Snodgrass, and R. H. Dauskardt. "Environmental Effects on Interfacial Adhesion." *Microelectron. Reliability* 41 (2001): 1615.
29. Lin, Y., et al. "Environmental Effects on Subcritical Delamination of Dielectric and Metal Films from Organosilicate Glass (osg) Thin Films." In *Materials, Technology and Reliability for Advanced Interconnects and Low-k Dielectrics—2003*, edited by A. J. McKerrow, J. Leu, O., Kraft, and T. Kikkawa, *Materials Research Society Symposium Proceedings*. Vol. 766, E9.4, Warrendale, PA, 2003.
30. Lin, Y., et al. "Subcritical Delamination of Dielectric and Metal Films from Low-*k* Organosilicate Glass (OSG) Thin Films in Buffered pH Solutions." In *Thin Films—Stresses and Mechanical Properties X*, edited by S. G. Corcoran, Y. -C. Joo, N. R. Moody, and Z. Suo, *Materials Research Society Symposium Proceedings*. Vol. 795, 93, Warrendale, PA, 2004.
31. Jacques, J.M., et al. "Fracture Property Improvements of a Nanoporous Thin Film via Post Deposition Bond Modifications." In *Materials, Technology and Reliability of Advanced Interconnects—2005*, edited by P. R. Besser, A. J. McKerrow, F. Iacopi, C. P. Wong, and J. J. Vlassak, *Materials Research Society Symposium Proceedings*. Vol. 863, B 3.8, Warrendale, PA, 2005.
32. Dersch, O., et al. "Diffusion of Water into Quartz and Silica Glass." *Mater. Sci. Forum* 248–49 (1997): 383.
33. Doremus, R. H. "Diffusion of Water in Silica Glass." *J. Mater. Res.* 10 (1995): 2379.
34. Shaw, T. M. "Moisture and Oxygen Uptake in Low-*k*/Copper Interconnect Structures." In *Advanced Metallization Conference 2003 (AMC 2003)*, edited by G. W. Ray, T. Smy, T. Ohta, and M. Tsujimura, 77, Warrendale, PA: Materials Research Society, 2004.
35. Xu, G., et al. "Moisture Diffusion along the TiN'SiO_2 Interface and in Plasma-Enhanced Chemical Vapor Deposited SiO_2." *J. Appl. Phys.* 88 (2000): 3695.
36. Tsui, T. Y., A. J. McKerrow, and J. J. Vlassak. "The Effect of Water Diffusion on the Adhesion of Organosilicate Glass Film Stacks." *J. Mech. Phys. Solids* 54 (2006): 887.
37. Tsui, T. Y., J. J. Vlassak, K. Taylor, A. J. McKerrow, and R. Kraft. "Effects of Absorption of Water and Other Reactive Species on the Fracture Properties of Organosilicate Glass Thin Films." In *Advanced Metallization Conference 2005 (AMC 2005)*, edited by S. H. Brongersma, T. C. Taylor, M. Tsujimura, and K. Masu, 695, Warrendale, PA: Materials Research Society, 2005.
38. Kingery, W. D., H. K. Bowen, and D. R. Uhlmann., *Introduction to Ceramics*. New York: Wiley, 1976.
39. Lin, Y., T. Y. Tsui, and J. J. Vlassak. "Water Diffusion and Fracture in Organosilicate Glass Film Stacks." *Acta Materialia*, 2007 (In press).

13

Chemical Vapor Deposition☆

13.1 Introduction: What Is CVD and Why CVD? **13**-1

13.2 Basic Aspects of CVD **13**-2
CVD Chemistry • Reaction Mechanisms: Thermally Activated Reaction, Plasma-Enhanced Reaction • ALD Reaction Mechanisms • Deposition Kinetics • ALD Process Characterization • Film Structures and Properties

13.3 CVD System Design **13**-16
Summary of Widely Used CVD Reactor/Systems • ALD Reactor Design: Next for CVD Equipment

13.4 CVD Thin Films **13**-22
Dielectrics • Conducting CVD Films

References **13**-82

Li-Qun Xia
Mei Chang
Applied Materials, Inc.

13.1 Introduction: What Is CVD and Why CVD?

Chemical vapor deposition (CVD) is a method of forming thin solid film on a substrate by the reaction of vapor phase chemicals which contain the required constituents. The reactant gases are activated by various energy forms such as chemical, thermal, plasma or photon, and reacted on and/or above the temperature-controlled surface to form the thin film. The reactive species, energy, rate of chemical supply, substrate temperature and substrate itself largely determine the film properties.

A wide variety of thin films are prepared by CVD for use in semiconductor device fabrication. The selection of materials such as polysilicon, silica glass, doped silica glass borophosphosilicate glass (BPSG), phosphosilicate glass (PSG), silicon nitride, tungsten, tungsten silicide, titanium nitride, and other emerging dielectrics, conductors, semiconductors is typically based on meeting integration requirements and cost target. The silica based materials chosen in the early semiconductor development is simply due to its compatibility with silicon in nature. The demands for shrinking geometry size and more functionality on devices put down more requirements: diffusion barrier (for sodium), step coverage by reflow or as-deposited conformality, film stress and interface control (adhesion, wetting). The never ever ended drives from consumers for lower price force manufacturers to balance between the cost of fabrication and performance enhancement. For some materials such as silicon, silica glass, and silicon nitride, and other dielectrics, CVD is the simplest and the most cost effective way. But for conductor materials, physical vapor deposition (PVD) is a more traditional way to deposit; only the cases where PVD cannot or very difficult to achieve, CVD prevail. One outstanding requirement is step coverage.

☆Li-Qun Xia (Dielectric); Mei Chang (Conductive CVD); Peter Lee (Low K dielectrics); Ian Latchford (Dielectric ARC); Pravin Narwanka (Ta_2O_5); Annabel Nickles (BST); Raman Achutaraman (Polysilicon); Hua Chung (ALD).

Tungsten, for example, required close to 100% step coverage to fill high aspect ratio via holes, is dominant by CVD. Some conductors such as titanium nitride and tungsten silicide are mixed depending on the applications and manufacturer's preference. Some applications such as aluminum fill; CVD Al and PVD Al have to work together in a cluster tool to accomplish the task.

Chemical vapor deposition reactors provide a controlled environment for the reactants activation, proper distribution, and delivery; in addition, the environment on and around the substrates. Successful CVD systems can provide not only the desired film properties but high throughput, reliable performance and low-operating cost. To enhance the overall system performance, many reactors design further incorporated in situ cleaning capability to maximize equipment uptime and minimize particulate generation; foreline exhaust and by-product management to reduce maintenance time; and/or integration capability with other sequential processes to reduce factory cycle time.

13.2 Basic Aspects of CVD

13.2.1 CVD Chemistry

The CVD films typically used in semiconductors include most common silicon-based materials: silica glass (SiO_2), doped silica glass (PSG, BPSG), fluorinated silicate glass, silicon nitride, silicon oxynitride, polysilicon, and doped polysilicon. Common metal CVD films include: tungsten and tungsten silicide, and titanium nitride are being well adapted in the fab. Carbon-doped porous silicate glass as low dielectric constant material has accepted in mass production. Chemical vapor deposition Al integrated W/PVD Al reflow for dynamic random access memory (DRAM) application also at the start of implementing into production. Additionally, high dielectric constant materials (hafnium silicate), and very low dielectric constant materials (carbon doped silicate glass) and copper barrier/seed (tantalum nitride, copper, ruthenium) are in development.

The number of potential chemistries leading to the commonly used films is huge, Table 13.1 lists only those chemistries that are or were widely used.

The chemistry played major roles in the resulting film properties and thus dictating its applications. As an example in the long history of silicon dioxide deposition, SiH_2Cl_2/N_2O high temperature oxide (HTO) deposited around 700°C could be used only around silicon substrate and gate polysilicon; SiH_4/O_2 low temperature oxide (LTO) was used over Aluminum metallization line. Tetraethylorthosilicate (TEOS) decomposition at high temperature can achieve equal or better film quality than HTO, but the cost prohibited its practical usage. Tetraethylorthosilicate/O_2 plasma deposition achieved much better step coverage than LTO and maintains the deposition temperature about 400°C, it replaced LTO in the application over aluminum interconnect in the late 1980s. All the evolutions were attributed to better chemistry. The game is still going on for almost all the films used in the industry.

One general rule found is most of CVD reactants are gases. Even for those liquid phase precursors, their vapor pressures are relatively high compared with other leading candidates. The vapor pressure defined the deliverability from bulk supply to substrate surface. During the course of delivery, the valve on/off, the flow rate control and monitoring, and the distribution, all the actions will induce flow restriction and pressure lost is unavoidable. High vapor pressure source is required.

For lower pressure chemicals, the shorter delivery line and/or larger diameter tube are preferred. To the extreme, point of use delivery by liquid evaporation could be desirable for being more stable and controllable. Tetraethylorthosilicate is the best example. Tetraethylorthosilicate is a very stable liquid; the silica deposition rate is directly controlled by the amount of TEOS delivered on to the substrate. Liquid TEOS metered by liquid flow controller are directly injected onto hot surface mounted in the gas panel right next to the reactor. All the silica deposition systems now equipped with direct liquid injection for all the liquid sources (TEOS, triethylphosphate (TEPO), and triethylborate (TEB)). For those semistable liquid precursors such as dimethyl aluminum hydride (DMAH), 1-methyl pyrolidine alane (MPA), most of applications still preferred bubbler for its easier replacement.

TABLE 13.1 Common CVD Deposition Chemistries

Film Type	Chemistry	Method	Application
SiO_2	SiH_4, O_2	Thermal	LTO passivation
	SiH_4, N_2O	PECVD	Intermetal dielectric
	TEOS, O_2	Thermal	Spacer
	TEOS, O_2	PECVD	Intermetal dielectric
	SiH_2Cl_2, N_2O	Thermal	Spacer
	TEOS, O_3	Thermal	Gap filling
	SiH_4, O_2	HDP CVD	Gap filling
BPSG	SiH_4, O_2, PH_3, B_2H_6	Thermal	Premetal dielectric
	TEOS, O_3, TEPO, TEB	Thermal	Premetal dielectric
FSG	SiH_4, SiF_4, O_2	HDP CVD	Intermetal dielectric
	SiH_4, N_2O, SiF_4	PECVD	Intermetal dielectric
SiN	SiH_4, NH_3 or N_2	PECVD	Passivation
		HDP CVD	Passivation
	SiH_2Cl_2, NH_3	Thermal	Oxidation or etch mask, spacer, etch stop
SiON	SiH_4, NH_3, N_2O	PECVD	Passivation
Polysilicon electrode	SiH_4, or Si_2H_6	Thermal	Resistor, gate, bit line, electrode
	SiH_4, PH_3, B_2H_6	Thermal	Gate, electrode
Tungsten silicide	WF_6, SiH_4	Thermal	Gate, bit line
	WF_6, SiH_2Cl_2	Thermal	Gate, bit line
Tungsten	WF_6, H_2	Thermal	Plug, local interconnect
Titanium nitride	$TiCl_4$, NH_3	Thermal	Electrode, adhesion layer
	TDMAT, N_2, H_2	Thermal	Barrier/adhesion layer
Aluminum	MPA, H_2	Thermal	Al plug wetting layer
	DMAH, H_2	Thermal	Al plug wetting layer

TEOS: TetraEthylOrthoSilicate; TEPO: TriEthylPhosphate; TEB: TriEthylBorate; TDMAT: TetrakisDiMethylAminoTitanium; DMAH: DiMethylAluminum Hydride; MPA: 1-Methyl Pyrrolidine Alane; PECVD: Plasma Enhanced CVD; HDP CVD: High Density Plasma CVD.

In order to deliver more precursors onto the substrate, one easy way is increase its vapor pressure by raising the temperature. The temperature has to be maintained all the way from the source to the reactor to prevent the precursor re-condensation. This is typically achieved by self regulated heat trace along the delivery lines or temperature controlled heat jacket. Cold spots could occur in unintentional miss on elbows, connectors, and induce liquid droplets in the delivery line, end up as particles or defects in the deposited films. WF_6 is a typical example. WF_6 has a boiling temperature of about 17°C, the fab operation temperature right around the temperature. Any actions on the delivery line such as fast evacuation, or large quantity usage of gases could induce condensation unless the line is well regulated and heated. The source bottle is actually preferred to be chilled to below room temperature, to assure any potential condensation will not occur in the line rather inside the source bottle. One could further increase the line temperature, since there is probability to hit the precursors stability limit of decomposition, especially for those metal organic compound targeted for more advanced applications. The decomposition reaction can set the limit for the temperature of gas delivery systems and could cause routine replacement of the delivery lines and valves due to its particulates contribution. Tetrakisdimethylaminotitanium (TDMAT), DMAH, MPA, TEB; these entire metal organic compounds exhibit the tendency.

Once the reactants are introduced on the substrate, the reactivity among reactants, between reactant and substrate determines the outcome. Some reactions are very fast and could be explosive such as SiH_4/O_2. It has to operate at low pressure such as low-pressure chemical vapor deposition (LPCVD) or high density plasma (HDP) deposition, to reduce the gas phase particle. Separated lines supply to the reactor until the point of application. Safety interlock is critical to prevent mixing in overflow situation. Some reactions are very slow and require external energy to activate such as SiH_4/NH_3 Plasma-enhanced chemical vapor deposition (PECVD) silicon nitride or run at elevated temperature such as SiH_2Cl_2/NH_3 and SiH_2Cl_2/N_2O. Some reactions generate "long" life time intermediate species and allow the

intermediates diffuse on the surface until it completes the reaction, thus enhancing step coverage performance such as TEOS/O_3 filling shallow trenches with silica and $TiCl_4/NH_3$ forming conformal titanium nitride liner and/or capacitor electrode. Some reactions can be very fast, but if one introduce the reactants one a time and react only on substrate surface, it can deliver very interesting results such as in Atomic layer deposition (ALD) processes.

The reaction byproduct should be volatile and easy to be purged, and in most of the cases, they are. One classical example is high temperature silicon nitride by SiH_2Cl_2/NH_3 chemistry, the reaction byproducts form solid adduct NH_4Cl. The solid adduct can cause defects on substrates and degrade the performance of pump, exhaust systems. The challenge is further expanded by the reactor dry cleaning. The cleaning chemistry typically is fluorine containing compounds such as C_2F_6/O_2 and NF_3, which would generate HF and other fluorinated byproducts which could be lingering in the process cavity and get incorporated inside the film and degrade the device performance. The halide by-product from reaction such as HCl from $TiCl_4/NH_3$ and from cleaning such as HF, SiF_4, can have unique affinity to NH_3 to form adducts like NH_4Cl, NH_4F, $(NH_4)_2SiF_6$ which are solid at room temperature and very difficult to eliminate. Exhaust line could be heated and the adducts would be pushed towards pump, collected after the pump station or one could have a trapping device to collect the adducts and/or to break NH_3 further down to harmless compounds N_2 by applying plasma power.

In addition, the reaction byproducts could react with the substrate materials and cause detrimental effects on microelectronic devices manufacturing such as delaminating or interface poisoning. The byproducts that were incorporated inside the films as impurities such as H, Cl, and F could induce gate threshold voltage shift, metal line corrosion and junction leakage. Halide chemistries in general, are not preferred. If halide chemistry is the only available choice, special precautions have to be paid for process optimization to minimize impurity concentration. For example, for depositing $TiCl_4/NH_3$ TiN at $>600°C$ to minimize its chlorine content, it is used in the area that could tolerate high temperature process such as memory contact formation and capacitor electrode. But in aluminum via contact or in more advanced contact application, low temperature without halides is a must; metal organic precursor for TiN becomes preferred. Another example is the tungsten silicide deposition, WF_6/silane chemistry is very simple and easy, but the film contains fair amount of fluorine and does influence on transistor threshold voltage; WF_6/dichlorosilane does get lower fluorine content, but pay the price of higher process temperature and more process complication. And PVD WSi_x does have its role in lowest impurity content and earned some process designers' preference. In addition, the uses of diffusion barrier layers are required to contain halogen contaminants for reacting with underlying materials to cause detrimental effects such as TiN that is required between Ti PVD film and W CVD film. Where W CVD is using WF_6 and H_2 for its superior step coverage, but the WF_6 and HF could readily attack Ti underlayer and delaminate the whole film stack.

13.2.1.1 Cleaning Chemistry

Most of the thin film materials used in the semiconductor industry can be cleaned or be reacted to form volatile byproducts to be pumped away from the reactors. Such as polysilicon, silicon oxide, doped polysilicon, doped silicon oxide, silicon nitride, tungsten silicide, and tungsten. These silicon and tungsten compounds can react with fluorine to form very volatile silicon fluoride and tungsten fluoride. Fluorocarbons and oxygen mixture plasma could generate plenty of free fluorine radicals and their reactions with those materials have been studied widely in plasma etching field. C_2F_6/O_2 is the first one to be implemented in CVD reactor clean in late 1970s. However, fluorocarbons specifically C_2F_6 and CF_4 have very strong green house effect in the earth atmosphere. The supply was restricted and severely reduced in late 1980s. NF_3 was then preferred.

NF_3 plasma in comparison with C_2F_6/O_2 plasma contains more free fluorine radicals and the cleaning rate is faster. But the NF_3 plasma is more aggressive on the reactor process kits, particularly the heater, where the highest temperature is located in the reactors. Another half of the story about the aggressiveness is due to the materials, aluminum, widely used in the plasma-enhanced CVD (PECVD) heater. Aluminum will react with fluorine to form non-passive aluminum fluoride on the surface, it can

sublimate easily above 400°C. It creates particulate inside reactor, degrades the heater surface texture, and eventually makes the heater lose its ability to control wafer temperature repeatedly. Remote NF_3 plasma decouples aggressive plasma from hot surface. And luckily enough, fluorine radicals could have a very good lifetime to be delivered handily into the CVD reactors when proper materials were used. Remote NF_3 plasma is now in the main stream.

Other chemicals such as HCl and ClF_3 also have been used in some cases. HCl is the cleaning gas for epitaxial silicon deposition for more than 30 years. The unique high temperature (about 1000°C) environment of epi reactor cannot tolerate the reactivity of fluorine, even Cl_2 is too aggressive. HCl reacts with silicon deposits to form silicon chlorides gases. ClF_3 is alternative supply of fluorine, it is even less aggressive than fluorine radical, but is more reactive than molecular F_2. ClF_3 is liquid in room temperature; it has the same constraint of handling F_2. The application is for tube furnaces nitride, and tungsten silicide chamber cleaning.

Waste emissions have got great public attentions as the volume of semiconductor fabrication has expanded rapidly. Greenhouse gases such as Freons, fluorine, and chlorine-containing gases used in CVD cleaning are subject to tightened regulations. C_2F_6 (Freon 116) transition to NF_3 has happened. The efficiency improvement of using NF_3 is one of the top of Continuous improvement program.

13.2.2 Reaction Mechanisms: Thermally Activated Reaction, Plasma-Enhanced Reaction

The CVD process can be generalized in a sequence of steps (Figure 13.1):

a. Reactants are introduced into reactor;
b. The gas species are activated and/or dissociated by mixing, heat, plasma or other means;
c. The reactive species are adsorbed on the substrate surface;
d. The adsorbed species undergo chemical reaction or react with other incoming species to form a solid film;
e. The reaction by-products are desorbed from the substrate surface;
f. The reaction by-product is removed from the reactor.

The most critical step is the chemical reaction on the surface to form the desirable film in step d. The rest of the steps are just to fulfill the materials transfer requirement. Although film growth is primarily accomplished by step d, overall growth rate is controlled by steps a–f in series, with the slowest step

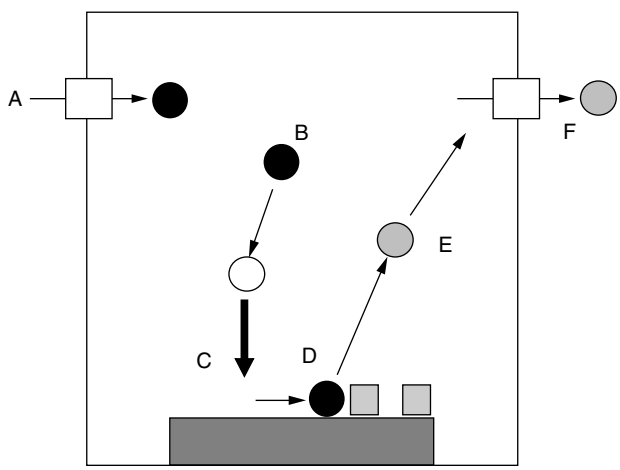


FIGURE 13.1 CVD reaction sequence.

determining the final growth rate. As in any typical chemical kinetics, the determining factors are the concentrations of surface species, wafer temperature, incoming charged species and their energies. The energy required to generate a chemical reaction includes kinetic energy, vibration energy as well as the chemical potential, e.g., F atom vs. F_2 molecule. Radio frequency (RF) energy tends to create active species with high chemical potential and high kinetic energy due to the external electric field.

Reactions may occur in the gas phase before reaching the substrate, or on the substrate surface among the reactants introduced and/or with byproducts; or on the reactor container walls (either hot or cold), or further down in the exhaust stream. Those non-productive reactions can effectively decrease the concentration and the rate of supply of desired active species and add potentially undesirable by-product species into the reaction. The non-productive reaction can also be detrimental by generating particles to induce defects on device, clogging foreline components, reducing the pump lifetime and effectively reduce systems productivity. The reaction inside the gas phase is usually referred to as a homogeneous reaction, which can be controlled by gas phase temperature and concentration dilution by inert gases. Cold wall reactors minimize competitive surface nucleation on the reactor wall. Therefore, the gas phase and surface reactions can be controlled separately to achieve desired film properties. Sometimes, a certain degree of gas phase reaction is desirable, if the reaction intermediates are required to form the final film.

The choice of reactants largely determines the extent of gas phase reactions vs. surface reactions. In some cases, such as SiH_4/O_2 (with or without PH_3, B_2H_6) or SiH_4/WF_6, the gas-phase reaction occurs as the gases mix together even at room temperature and deposition will occur on any surface they encounter. To control the reaction, system designers concentrate on distributing and delivering the gases directly to the wafer surface with minimum interference. Low-pressure operating conditions are preferred to reduce premature gas phase reactions. In other cases, alternative source gases or liquids are selected to avoid gas-phase nucleation or improve step coverage; these include TEOS, $TEOS/O_2$, TEB, TEPO, SiH_2Cl_2/WF_6, WF_6/H_2.

To minimize undesired deposition in the reactor, single-wafer or batch reactors with limited heated area, called "cold wall reactors," have gained acceptance. In the case of thermal deposition, any hot surface in the path of reactants flow will foster deposition. The cold wall reactor design provides the conditions for idealized gas delivery and maximized gas utilization for better control of film properties. However, energy efficiency is lost due to greater cold surface area which takes away heat.

In general, higher deposition temperature drives out impurities more efficiently, and the deposited films have higher density and a more crystalline structure. In the early years of semiconductor manufacturing, high temperature ($>600°C$) films were preferred. High-temperature films are still used around the gate and transistor area. However, overall thermal budget reduction and low-temperature multi-level metallization drive the developments of lower temperature processes, including new chemistries and plasma configurations.

Plasma-enhanced deposition typically uses less reactive gases such as SiH_4/N_2O, $TEOS/O_2$, and SiH_4/NH_3 to avoid gas-phase reactions. The plasma dissociates the precursor and creates high-energy forms of the reactant species that accelerate the reaction rate at much lower temperature than without the assistance of plasma. Since the generation of active species is directly tied to the plasma, the power input and flow rate of SiH_4 or TEOS dictate the deposition rate, while temperature has less effect.

Associated with plasma are charged species such as ions and electrons. The substrate surface not only receives active precursors but is subject to the bombardment of charged species. The energy of charged species depends on cathode, anode, ground geometry and RF frequency, waveform, which together are called the *RF configuration*. The short-lived active species react and deposit on the surface, the thermal energy and ion bombardment continue to modify the deposited materials. Process temperature controls the surface and bulk diffusion of active species, while the RF configuration controls the active species including ions distribution and their energies distribution; both together influence film properties: structure, morphology, density, stress, and impurities. The plasma-enhanced deposited films tend to be of smaller grain size, or even amorphous, and contain certain amounts of impurities such as hydrogen, carbon or halide atoms.

The combination of low temperature, self-cleaning capability and versatile film tunability has assured the position of PE CVD in the semiconductor industry. To minimize deposits on the reactor surfaces, limiting the plasma area is beneficial. The standard parallel plate configuration provides an efficient design to focus the bulk of deposition on the wafer. At the same time, the reactor's plasma capability also provides the potential for in situ plasma cleaning by introducing etchant cleaning gases such as C_2F_6 or NF_3 to remove silicon dioxide and silicon nitride deposition from chamber surfaces. One limitation of plasma deposition involves the potential charge imbedded in the film and an uneven charge effect on the finished device. For this reason, the closer to the transistor structure, the more reluctant chipmakers have been to use plasma-enhanced deposition. Thus, the plasma enhanced deposition was introduced from backend passivation such as silicon nitride, silicon oxynitride, phosphorus doped silica; to intermetal dielectrics such as plasma TEOS, HDP silica.

To overcome the concern of charge damage and still maintain the advantage of low temperature processes, two approaches are pursued. One is remote plasma instead of in situ plasma. Reactants are plasma dissociated or activated remotely, then introduced onto the substrate surface along with second reactants to complete the reaction. The reaction mechanisms are very similar to thermally driven processes. But one has to consider the short life time of the activated species and how to distribute over the large substrate surface. There is only one close related successful example, TEOS/O_3. The O_3 is very reactive and easy to decompose back to O_2. Point of use generation of O_3 by corona discharge at high pressure ($>$1 atm) has to be part of CVD system design. Fortunately, the O_3 is stable enough and concentration could be high enough to produce reasonable silica deposition rate and provide good step coverage. All other trials become academic and facing significant challenge in the resulting film's uniformity, a trade-off between active species life time and distribution.

13.2.3 ALD Reaction Mechanisms

Another approach is ALD, which has much more promise by modifying the reactant delivery sequence. The basic sequence is alternating A cycle and B cycle.

A cycle:

- Reactant A is introduced into reactor;
- Reactant adsorbed on substrate surface and react with surface species to form a solid layer;
- The remaining gas species are pumped and/or purged away.

B cycle:

- Just replace reactant A by reactant B in A cycle.

Repeat A cycle and B cycle until the film reaches the desired thickness. Where the reactant A or B could be plasma or remotely enhanced, one could modify the sequence to include C cycle with third reactant C or more cycle with other reactant, and rearrange the sequence to reach desirable film stack.

In an ideal case, one monolayer is deposited in each cycle. The film thickness can be precisely controlled by number of deposition cycles. In reality, the film thickness obtained in one cycle can be varied from 0.1 to 5 Å depending on the chemistry and the materials deposited. The advantage of ALD is its thickness control, an inherited control by finite thickness in each deposition cycle. Within each cycle the reactants have sufficient time to diffuse the all available surfaces and to complete the reaction. An ideal ALD should automatically provide excellent step coverage, uniformity, and repeatability; these are critical for ultra thin film applications such as capacitor dielectric, capacitor electrode, Cu barrier, adhesion layer, and gate high k dielectrics.

For ALD process, the initial film growth strongly depends on the presence of surface reactive function groups. To illustrate the actual reactions taken place in each exposure step, the growth mechanism of Al_2O_3 using Trimethyl aluminum (TMA) and H_2O is shown in Figure 13.2. If the substrate surface is hydroxylated prior to deposition, TMA molecules are ready to react with the surface functional (hydroxyl) groups. The reaction involves the releasing of CH_3 ligands from TMA and the molecule is firmly bound

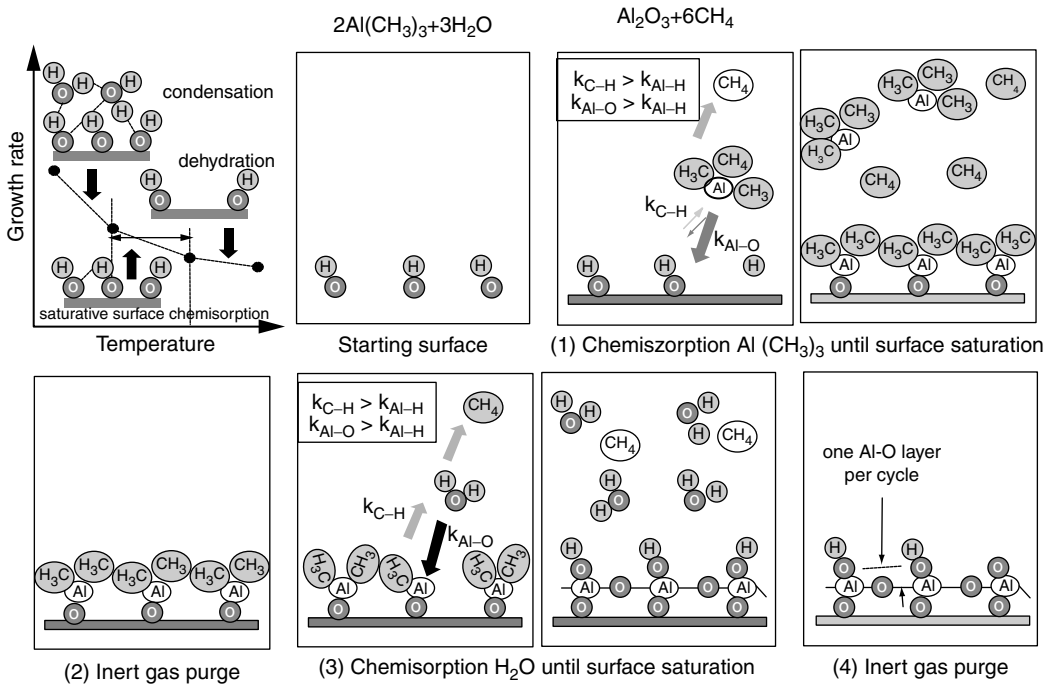


FIGURE 13.2 Schematic diagram showing atomic layer deposition Al_2O_3 growth mechanism by TMA and H_2O.

to the surface by forming Al–O bonds. The reaction byproducts, CH_4, and excessive TMA molecules are evacuated in the following purge step. In the next step, the substrate is exposed to H_2O doses. H_2O molecules react with the surface groups Al–$(CH_3)_2$ and lift CH_3 ligands forming CH_4 as reaction byproducts. After H_2O exposure, the surface is terminated by hydroxyl groups, i.e., converted back to its original state and ready to react with TMA in the next deposition cycle. Excessive H_2O and CH_4 molecules are removed during the subsequent purge step. A monolayer of Al_2O_3 is formed on substrate surface after one deposition cycle. The growth process is repeated in cyclic manner until a desired thickness is reached.

13.2.4 Deposition Kinetics

On the substrate surface, the detailed chemical pathway to the solid film is typically very complicated. The study of thin film deposition kinetics can be roughly categorized into incoming species control and surface reaction control, since they are sequential events. The deposition rate of incoming species can be determined by the supply (flow rate and/or RF power) or the diffusion rate through the flow and/or plasma's boundary layer, these factors are not very sensitive to temperature. The surface reaction, however, typically has strong temperature dependence.

At lower temperatures and particularly in thermally activated reactions, gas phase and surface reactions are much slower than the rate of incoming species; thus the deposition rate is limited by the kinetic reaction rate, most of time it is referred as "surface reaction controlled regime," and more generally called "kinetically limited regime." At higher temperatures or in most plasma-enhanced reactions, the incoming species can be the limiting factor, and is called a "mass transport limited regime" or "supply limited regime" or more broadly "incoming rate control regime."

In the case of chemistries without gas phase reactions and purely thermal driven reactions such as TEOS thermal decomposition, SiH_4 thermal decomposition, WF_6/H_2 thermal reaction or DMAH

thermal decomposition, these two regimes can be easily defined. But for other cases such as SiH_2Cl_2/WF_6, $TEOS/O_3$ or any plasma-enhanced deposition where active species are generated through complicated routes before reaching the substrate surface, distinguishing between the two regimes becomes difficult or it may not exist at all.

The incoming rate control regime can further be distinguished by its sensitivity to reactants' raw supply rate or gas distribution configuration, and by RF power in the plasma-enhanced deposition case. If the reactant supply is the dominant factor in controlling deposition rate, it is truly "supply limited," or the so-called "starvation regime." If the gas distribution itself is sensitive to the deposition rate, typically occurring at high-pressure and high-temperature processes such as atmospheric pressure epitaxial Si deposition, the flow or thermal boundary layer material diffusion limits the deposition process and is truly a mass transport limited regime. In both cases, the reactions on the substrate surface typically have sufficient time to be completed and it results in better purity and better crystalline structure. In plasma-enhanced deposition, typically there is no obvious flow or thermal boundary layer due to lower operating pressure and lower temperature, thus no "mass transport limit" case; but there is transition between supply limited regime to "plasma power limited regime." At low supply flow rates, plasma power can completely dissociate the incoming supply, the deposition rate is proportional to incoming flow rate, and the film contains less impurities; at high supply flow rates, the plasma power cannot keep up with the supply, the deposition rate will strongly depend on the plasma power, and the film contains more impurities.

In many practical CVD cases, in their normal operation conditions, the feed rate of one of the gases is the primary factor determining the deposition rate; i.e., the film deposition rate is proportional to the flow rate of a reactant gas, such as SiH_4 in SiH_4/O_2 silica or plasma SiH_4/N_2O silica, TEOS in plasma $TEOS/O_2$ silica, or WF_6 in SiH_4/WF_6 tungsten silicide. These are clearly operated in the incoming rate control regime. But the flow rate linear dependency is not an absolute criterion for determining the regime the process was tuned in.

There are more practical cases operated in kinetically controlled regime: polysilicon deposition by SiH_4 thermal decomposition, silica deposition by TEOS thermal decomposition and by $TEOS/O_3$ thermal reaction, tungsten deposition by WF_6/H_2 thermal reaction, titanium nitride deposition by $TiCl_4/NH_3$ thermal reaction, and aluminum deposition by DMAH thermal decomposition. In these cases right in surface reaction limited regime, the deposition rates also depend on the flow rate of the reactant. The surface reaction can be a linear function of some absorbents which is also a linear function (in short range) of partial pressure of the feed gas such as DMAH in thermal decomposition. When the flow rate of DMAH is increased while keeping the same dilution flow, the DMAH partial pressure is increased which in turn increases deposition rate. Similar cases also have been observed in other metal-organic precursors where adsorption and desorption on the surface play significant roles in the rate limiting steps. The competition at the adsorption site on the surface between incoming precursors and reaction by-products make the rate-limiting equation a little more complicated. The best way to determine the regime was the processes were operated by examining the deposition rate and step coverage change through a broad temperature range and flow range. And lots of processes were tuned right in the transition area to balance the film properties and step coverage.

One way to break the mutual dependency between film properties and step coverage is by ALD. Atomic layer deposition separated the incoming rate and the surface reaction rate. The feed rate can always be saturated by exposure time and dosage, and surface reaction could always be completed by increasing the time of adsorption and reaction time in step b.

13.2.5 ALD Process Characterization

In this section, the most commonly used approaches to characterize ALD processes are discussed. The major steps include reaction gas flow rates, reactant exposure time, inert gas purge time, and deposition temperatures. In these experiments, the film growth rate/cycle is measured by varying each parameter independently. On the other hand, many factors such as substrate surface conditions, precursor

self-decomposition, insufficient purging, impurity incorporation and byproduct re-absorption blocking reaction sites can result in deviations from ideal ALD processes and cause difficulties to interpret the process results. Examination of film properties can provide certain useful information to understand the growth mechanisms. However, in situ surface chemistry analysis is required to characterize the react mechanism of an ALD process in great details.

13.2.5.1 Precursor Exposure Time Characterization

Atomic layer deposition is a self-limiting reaction process. There is no further film growth once the surface is completely saturated. On the other hand, insufficient exposures can result in severe impacts to the film quality such as poor thickness uniformity, high impurity incorporation, and poor film conformality. It is critical to make sure saturation exposure is reached.

The experiment is carried by varying the duration of reaction gas exposure, one at a time, and keeps the other parameters constant. As the exposure time is increased, more chemisorption sites are covered by incoming precursor molecules until saturation is reached. It is expected that the length of exposure time can be varied depending on the reactivity of incoming gas molecules with surface function groups and precursor flux. A commonly used unit to measure precursor dosage is langmuirs (L), which is the product of precursor partial pressure and exposure time; $1\ \mathrm{L} = 1 \times 10^{-6}$ Torr-s. The required amount of precursor dosage is mainly determined by its reactivity and surface areas to be covered. For metal and metal nitride ALD processes, metal precursor exposures are typically less than 1×10^5 L. In contrast, much larger doses can be required for reducing agent exposures ($>1 \times 10^7$ L) if the reducing reaction is slow. At the first approximation, it is expected that the exposure time can be reduced by increasing precursor fluxes. In some cases, however, the growth rate may increase only with exposure time instead of precursor fluxes if the surface reaction is very slow.

Once the surface is saturated, there is no further growth expected. However, if the precursor can be thermally decomposed at the process temperature, the growth process can be continued by further increasing the exposure time and the slope of growth rate vs. exposure time will depend on precursor decomposition rates. If the decomposition rate is much slower than surface reaction rates, the initial growth is dominated by surface reaction process. A very steep slope will be observed in this stage. The decomposition process become dominating only after surface reaction is completed and the slope is determined by precursor decomposition rates.

13.2.5.2 Purge Time Optimization

In ALD, inert gas purging is required to separate different reactions gases. Insufficient purge can cause overlapping between reactants leading to parasite CVD reactions. As a result, the growth rate and thickness non-uniformity are increased. The onset of CVD reaction can result in poor film step coverage, increasing surface roughness and impurity concentration. In some cases, the growth rate decreases with increased purge time, even reaction gases are well separated. This happens if precursors are physically absorbed or condensed on substrate surfaces during the course of precursor exposure. Since the surface species are not stable, the surface species can be desorbed during purge steps and a lower growth rate can be obtained.

13.2.5.3 Temperature

Multiple growth mechanisms can occur by varying ALD process temperatures. Based on the characteristics of growth processes, the growth temperature can be separated into three zones.

13.2.5.3.1 Low Temperature

- Growth rates linearly increase with temperatures.
  This indicates that the reaction rate is slow and surface reaction is not completed. The steepness of the slope can be changed by varying the dosage of the reducing agent. The larger the dosage is, the steeper the slope can be obtained.

- Growth rates linearly decrease with raised temperature.
  This happens when precursor condensation occurs. Multiple layers are absorbed on the surface.

13.2.5.3.2 Medium Temperature

In this temperature range, the growth rate to deposition temperature dependency is a minimum. In some cases, the growth rate is independent from growth temperature. This can be observed only when the growth process is self-limited and the density of chemisorbed species is temperature independent. In most cases, minor temperate dependencies are observed. A positive slope is observed when slow precursor thermal decomposition process occurs. Negative slopes can be observed if the density of chemisorbed species is temperature dependent and sticking coefficient is reduced with increased temperatures. Atomic layer deposition processes should be performed in the temperature range to obtain maximum process windows.

13.2.5.3.3 High Temperature

- Growth rate linearly increases with raised temperature.
  The slope of growth rate vs. temperature is much steeper than that in the medium temperature range because of rapid increases of precursor decomposition rates at high temperature.
- Growth rate linearly decreases with raised temperature.
  The slope of growth rate vs. temperature can be much negative than that in the medium temperature range because the surface sticking coefficient of precursor molecules decreases rapidly with raised temperatures. The density of chemisorbed species is lowered with increased temperature.

Figure 13.3 summarizes different kinds of growth rate vs. temperature curves which can be observed in ALD processes. The growth temperature should be chosen in the middle range where the growth proceeds in the self-limiting manner.

An ideal ALD process proceeds with a self-limiting growth mechanism. In reality, many non-idealities can happen. The frequently observed deviations include incomplete reactions, insufficient purge time and precursor decomposition. Certain deviations could be acceptable without losing the advantageous characteristics of ALD if the non-ideal conditions proceed in surface controlled mode rather than mass transport controlled modes. These non-idealities can be resolved by

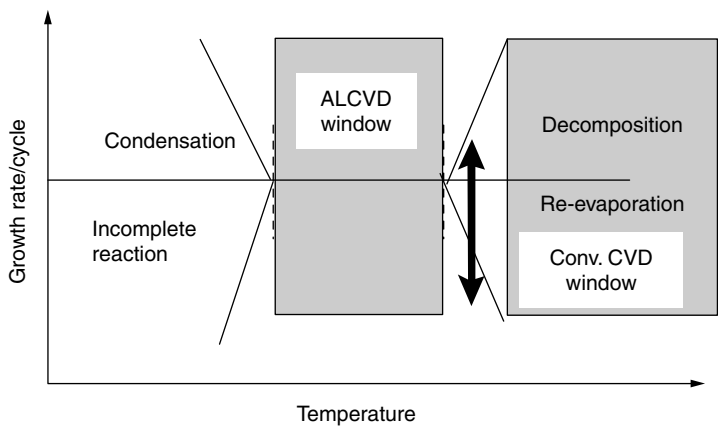


FIGURE 13.3 Growth rate vs. deposition temperature curves.

- Increase the exposure time to ensure the surface is saturated;
- Increase purge time to avoid CVD reactions;
- Reduce the deposition temperature to avoid precursor decomposition.

One should keep in mind that these changes may associate with negative effects. For example, increasing purge or exposure time will impact on throughputs if a system is used for production. Reducing process temperature may result in film quality degradation if reactions are not completed. The process should be optimized to obtain the best film quality without significantly scarifying the productivity.

Atomic layer deposition processes provide benefits over CVD processes such as

- Large temperature windows;
- Low impurity contamination;
- Low risk of particle generation by gas phase reaction;
- Easy to scale up to large wafer size;
- Relax the requirement of precursor vapor delivery system and chamber hardware design;
- Improve precursor utilization;
- Precise thickness control;
- Excellent film conformality;
- Flexibility on the tailor film composition;
- Capability to deposit nanolaminate composite layers.

Although ALD provide many advantages over CVD, the low growth rate of ALD process strictly limits the technology for ultra-thin film deposition. In addition, ALD processes are more prone to surface conditions. When the ALD sequence is in steady state, i.e., alternating reactant A cycle and B cycle, after ending each cycle the surface was well prepared for the following cycle. But the fresh substrate surface may not be suitable for initial absorption or reaction. In this case, incubation time or delay nucleation will be observed. Surface treatment to modify the surface function group and/or extend the reactant soaking time is required.

The requirements for microelectronic devices have increased throughout their whole history and this has put high demands on microelectronic manufacturing. Transistors and other microelectronic components have to become continuously smaller and faster. Atomic layer deposition becomes one of the key enabling technologies to meet stringent process requirements in semiconductor microelectronic device manufacturing. The major applications are

- High-*k* dielectric and metal gate for front end applications;
- Liner, barrier and seed layer deposition for metal processing;
- Deep trench electrode for DRAM manufacturing surface sensitivity.

In both CVD and ALD cases, when reaction rate could be limited by surface reactions or surface adsorption, desorption, dependency of substrate surface will show up. Surface sensitive behaviors give rise to beneficial deposition properties or create headache for process engineers. Terms such as "selective deposition," "incubation time," "nucleation" and "preferential growth" describe these surface-sensitive behaviors. In these cases, the deposition is easier or faster on one surface and retarded or slower on another. Selective tungsten is a typical example; at low temperature (300°C–350°C), H_2/WF_6 will be selectively deposited on Si or other clean metal surface relative to silicon dioxide. The selective tungsten was explored extensively to fill contact and/or via in the late 1980s But it failed due to incomplete control of surface reaction and chemistry incompatible with substrate. Even in blanket tungsten case using the same H_2/WF_6 chemistry, the nucleation of tungsten on TiN substrate is accomplished by different chemistry such as SiH_4/WF_6. This only indicates H_2/WF_6 chemistry inherently depends on surface in this temperature range. Other example is Al deposition on via on SiO_2 by DMAH. Dimethyl aluminum hydride would not easily decompose and nucleate Al on silicon dioxide surface, but it is ready deposit Al on TDMAT flash treated surface. Tetraethylorthosilicate/O_3 also showed the deposition rate difference on

Al patterns verses on SiO_2 surface. And there are cases in ALD. All the examples are thermally driven processes, plasma enhanced processes have much less dependency on substrate surface. Since the device wafer surface has already gone through many process steps and often been exposed to ambient prior the CVD process, it cannot maintain absolute control of surface impurities (C, O, H) and atomic level defects. The control of selectivity become an art and difficult in the production environment. In the practical applications, in situ surface treatment or a separated nucleation process is normally implemented to eliminate the surface sensitivity instead of trying to control the selectivity. The surface treatment will become much more critical when the required film thickness decreases and branch into a dedicated subject of interface engineering.

13.2.6 Film Structures and Properties

The ideal film properties and structures are dependent on their specific applications. For example, in passivation, the films must be moisture and Na diffusion barrier with good step coverage and no pinholes. Silicon nitride is an ideal material for the passivation layer, but high-temperature thermal deposited nitride exceeds the temperature requirement for Al metallization. Plasma-enhanced nitride film can satisfy the temperature requirement, but the H content is a concern because hydrogen can cause hot carrier lifetime degradation. Controlling the hydrogen content by chemistry, by various RF configurations, and by operating temperature have been topics of interest in passivation. In intermetal dielectric application, filling between aluminum lines without void is required. Plasma TEOS deposited silicon dioxide exhibits very decent step coverage, but not good enough to completely eliminate the void without sputtering to remove overhang. When line pitch is getting smaller, the gap aspect ratio increased, plasma TEOS together with sputter etch ran out of steam; it was replaced by HDP deposited silicon dioxide with built-in sputtering capability. The emergence of copper metallization and the implementation of copper chemical-mechanical polishing eliminated the requirement of filling high aspect ratio gap. The old silicon dioxide deposition technologies then revived. In the course of changing filling capability requirement, other properties such as dielectric strength, low stress, cracking resistance, good adhesion, no fix charges, no mobile ions, low hydrogen mobility, all have to meet their requirements. Chemical vapor deposition process parameters have to be fine adjusted to meet all the film properties requirements, in addition, the production requirements.

13.2.6.1 Composition

Thermal processes typically are very difficult to alter the film compositions in terms of main ingredients such as Si/O ratio in silica deposition, Ti/N ratio in Titanium chloride/ammonia TiN deposition. There are exceptions however; W/Si ratio can be controlled by WF_6/SiH_4 or WF_6/SiH_2Cl_2 ratio due to the silicon source gas decomposition in competition with the chemical reaction between fluoride and hydride. On the other hand, plasma process can easily tune the composition by varying the incoming gas ratio such as SiH_4/N_2O plasma silicon dioxide, and SiH_4/NH_3 plasma silicon nitride.

Dopants such as boron, phosphorus in silicon dioxide can greatly lower its viscosity and allow it to reflow less than 800°C. The phosphorus can further getter sodium to prevent its penetration to sensitive gate area. The control of the dopants typically is through the dopant source flow such as B_2H_6, PH_3, TEPO, and TEB which depends on the deposition chemistry. The efficiency of dopant incorporation is controlled by the mechanism of dopant sources decomposition. In thermal process, temperature is the dominant factor. In plasma enhanced process, the temperature dependence is much less, plasma power is much more critical. The same tuning principles were applied on polysilicon deposition process too.

Concerning on the impurities content, similar parameters such as reactants flow ratio, plasma power, and substrate temperature are the most important. Hydrogen content in silica in SiH_4/O_2 and SiH_4/N_2O plasma chemistry, hydrogen content in silicon nitride in SiH_4/NH_3 plasma chemistry and carbon content in $TEOS/O_2$ plasma and $TEOS/O_3$ chemistry, and fluorine content in WF_6/SiH_4 tungsten silicide deposition are typical examples. To reduce them, one can just increase the deposition temperature until it hits the desirable temperature limitation; lower the main source supply flow of SiH_4 or TEOS; increase

the consumption rate by increasing the companion reactants flow such as O_2, O_3, NH_3, etc. High operation pressure effectively increases the residence time of the reactants and byproducts; and it tends to retain more impurities in the films. For plasma enhanced processes, one could increase the RF power to enhance the completeness of the reactant dissociation. RF power also controls ion bombardment energy to densify the film and to reduce light elements content by preferentially sputtering.

13.2.6.2 Microstructures

Microstructure of CVD film is not simply controlled by process parameters. There are several convoluted competing mechanisms including film growth rate, nucleation rates on substrate surface, and diffusion-controlled grain growth rate in the film itself. During the nucleation stage, the grains can line up with the substrate if a surface texture is available; or the grains could nucleate in random orientation. In examples such as CVD W on sputtered W or CVD Al on sputtered Ti, the preferred orientation will follow the underlayer crystal structure and typically very smooth and reflective. However, if the substrate is randomly oriented grain structure or amorphous structure, such as CVD Al, Cu, or metalorganic TiN on SiO_2, the films tend to be randomly oriented. In addition, the nuclei would form islands instead of continuous film when the materials Al and Cu have high surface mobility, and the resulted films are very rough. The film deposition growth mechanism eventually takes over the control of film structure after passing the nucleation stage; for example, $TiCl_4/NH_3$ thermal deposited TiN and CVD polysilicon in certain operating ranges show strong columnar structure on silicon dioxide substrates, although the nuclei grow in random orientation. The fast growth orientation eventually outgrows the other orientation and becomes dominant in the film structure. The grain structure could be evolving after its initial formation. In the case of polysilicon deposition, the silicon grain growth mechanism could run parallel during the film growth and result in much larger grains or rough surface if the deposition rate is slow compared to the grain growth rate. But in TiN case, the grain growth rate is much slower in the CVD deposition conditions (600°C); it always grows in columnar structure. In the case of tungsten deposition in WF_6/H_2 chemistry, the grain growth rate also is very slow in the deposition temperature (350°C–450°C); the microstructure actually depends on the deposition regime. In supply limited regime, tungsten grew in nice, smooth columnar ⟨111⟩ orientation. But in the surface reaction regime, the tungsten grains grows larger and rougher with film thickness and show some ⟨110⟩ preferred orientation, although the tungsten nucleation film deposited in WF_6/SiH_4 chemistry is in random orientation. Another example is ALD TaN grown in copper dual Damascenes structure. Atomic layer deposition TaN process allows amorphous film to grow inherently since the deposition temperature is too low to allow atomic diffusion to occur. But if the TaN is grown on crystalline Cu, the TaN lattice will line up with Cu lattice in the initial couple of layers like epitaxial growth. When the film grows thicker, it transits back to amorphous.

For amorphous films such as silicon dioxide and plasma silicon nitride, porosity and density are the main film structure characteristics of interest. Those properties are affected by gas-phase reaction, reaction chemistry and ion bombardment. Gas-phase reactions will form powdery, granular, and rough surfaces and exhibit high porosity and low density, such as in the case of SiH_4/O_2 LTO. The effect of ion bombardment is shown in the case of PECVD SiN: higher RF power deposited film always has denser structure and has compressive stress while the nitride film at low RF power and low temperature could be rough and porous. High density plasma deposited silicon dioxide also demonstrated the ion bombardment at low operating pressure that overcomes the nature of gas phase reaction of SiH_4/O_2 and produced good and dense film. The chemistry plays a subtle role. One could compare the plasma silicon nitride properties deposited by the chemistries of silane/ammonia and of silane/nitrogen. The silane/ammonia chemistry, silicon nitride film typically has better step coverage, better side wall film integrity and overall less pinhole density. The hydrogen rich chemistry could have been enhancing the surface mobility or the lifetime of intermediate species. Hydrogen could play a similar role in high density, plasma silicon dioxide deposition process by silane/oxygen chemistry; the addition of hydrogen could improve the gap filling capability.

In some applications, post anneal is required; the microstructure of the deposited films will be modified and/or phase transformed. Tungsten silicide typically deposits over doped polysilicon and forms a polycide stack. Since deposited tungsten silicide is amorphous, it is deposited by silane/tungsten hexafluoride chemistry or hexagonal phase, if it is deposited by dichlorosilane/tungsten hexafluoride chemistry. Annealing about 900°C transforms the crystal structures into more stable tetragonal phase. Any excess silicon in the deposited silicide films will be redistributed at the polysilicon/silicide interface and the resulted tungsten silicide films silicon/tungsten ratio typically will be about 2.2 regardless the initial compositions.

13.2.6.3 Stress

When the films stack together, as in the multilevel interconnect devices, the various films' mechanical stress could build up high enough to crack the films, extruding the embedded metal lines to form void and destroy the devices functionality. In the transition to low dielectric constant, copper interconnect, the stress have shown much more subtle impact compared with aluminum silica interconnect. On one hand, low *k* dielectrics have much higher tensile stress and less yield strength vs. conventional silicon dioxide. On the other hand, copper is much easier subject to stress migration. Film stress integration becomes much more important. In the transistor gate area, the film stress has been demonstrated and utilized to promote the electron and hole mobilities in the transistor channel.

Typically, any thermal driven CVD processes produce tensile film. Two factors control the final stress: deposition as grown stress and the thermal expansion coefficient difference between the film and silicon substrate from deposition temperature to room temperature. Silicon dioxide has lower thermal expansion coefficient than silicon. Assuming the deposited stress at temperature as zero, the resulting room temperature stress should be slightly compressive. But if the deposition process did not generate fully densified film, the as grown stress is tensile; the room temperature stress could be tensile. After annealing at high temperature, the stress could be relaxed; the resulted stress could be back in compressive. Other materials, all have higher thermal expansion coefficient than silicon, all exhibit tensile stress; the more the refractory metal or compounds, the higher is the tensile stress. Tungsten, tungsten silicide, titanium nitride, and silicon nitride all exhibit high 10–20 Gpa stress due to the combination of as grown stress and thermal expansion induced stress. Process operation regime such as supply limited regime, could offer lower impurity content and better crystal structure and thus slightly lower stress. But this regime did not offer any practical application for good step coverage. In soft metals such as aluminum and copper, the as grown stresses are typically very low and stresses can also be relaxed through plastic deformation.

To tune the film stress, external force is a must. Plasma CVD processes, with the power of ion bombardment, typically shows compressive stress and stress that can be tunable by process conditions. The RF power is the biggest knob. The ions generated in the plasma, accelerated through the plasma sheath on the substrate, physically compact the deposited film. RF configuration with high anode to cathode area ratio induces high cathode potential, or self bias can booster up the ions energy to compact the film much easily and even to move materials around as in the case of HDP deposition.

13.2.6.4 Step Coverage

Step coverage is one of the main advantages of the CVD method, especially when comparing metal CVD to PVD. To get good step coverage, the inherent chemistries and operating conditions are critical. For example: the high reactivity of SiH_4 with other reactants (O_2, WF_6, and NH_3) inhibits good step coverage, but by itself, SiH_4 decomposition allows excellent step coverage. Once the chemistry is tuned to permit a gentler reaction, the temperature and reactants supply rate can be balanced in the surface reaction limited regime to achieve excellent step coverage, even filling high aspect ratio holes, as seen in the chemistries of TEOS, TEOS/O_3, WF_6/H_2, and some metalorganic compounds. One interpretation of this good step coverage is that the reactant species have excellent mobility with enough, long life time on the surface to migrate before they decompose or further react. One exception to this model is HDP oxide deposition, in which the step coverage is achieved by physical ion bombardment knocking off step

corners and redistributing the incoming materials to the side wall. The extreme case is ALD; the controllable long delivery time without reaction allows the precursors to reach much further into the feature before the next cycle of reactant comes to complete the reaction. The step coverage can be as good as to penetrate porous materials deep into the materials structure.

13.2.6.5 Interface Properties

The bonds between substrate and film have to stand for the full device processes integration, in addition to the films themselves. The nature of chemical bonds determined the strength. Silicon to silicon oxide, aluminum to aluminum oxide and aluminum oxide to silicon oxide all have very good adhesion strength. The multilevel interconnect devices built from silicon dioxide dielectrics and aluminum metal lines on top of silicon substrates are inherently stable and strong. Chemical vapor deposition tungsten does not have good adhesion on silicon oxides, titanium nitride as adhesion layer was introduced to hold the tungsten plugs inside the silicon oxide matrix. Although the materials selections are designed to be chemically compatible; CVD processes themselves could modulate the interface properties by the way of CVD processes initiated. In the case of tungsten and tungsten silicide deposition, tungsten hexafluoride has very high reactivity with substrate materials titanium nitride and polysilicon, the deposition processes have to introduce silane before tungsten hexafluoride. The interface tends to be slightly silicon rich to preserve the adhesion property. Another example is silicon nitride barrier deposited on copper damascene structure. The silicon source gas silane can easily react with copper at 300°C to form copper silicide. The silane has to be introduced after ammonia and after a light passivation layer on copper formed; in addition, a spike of silane gas at the introduction has to be eliminated. The nucleation process or the initiation process is often a significant part of CVD processes recipes development.

The final top surface after the bulk CVD film deposition could have some concerns too. The process recipe sequence of terminating the gas streams and/or plasma could result in the film sensitivity to ambient exposure, surface hydrophobic or hydrophilic property, and surface defects. These do not have direct impact on CVD film properties themselves, but influence the device processes integration.

13.3 CVD System Design

The CVD reactor designs are centered on substrate, accommodated with individual forms of energy supply and gas delivery for each chemistry and application. The CVD system include reactors, gas delivery systems, exhaust systems and wafer handling system, controller, and software. They can be categorized along several lines:

1. Operational pressure: atmospheric and sub-atmospheric (SA), reduced pressure, and low pressure;
2. Reaction energy input: thermal and/or plasma enhanced;
3. Substrate energy input form: radiant heat, induction heat, resistance heat;
4. Reactor wall temperature: hot wall, cold wall;
5. Number of wafers in reactor: single-wafer, batch, and continuous batch.

For PE CVD systems, further detailed category could be based on the excitation frequency: low frequency (<1 MHz, typical at 350–400 KHz), high frequency (>1 MHz, typical at 13.56 MHz), and dual frequency (high and low frequency); and on RF configuration: parallel plate, reactive ion etching (smaller cathode), inductively coupled plasma, microwave, and electron-cyclotron resonance (ECR).

Chemical vapor deposition systems evolution originated from the laboratory experimental reactors and their gas delivery methods aiming to deposit thin films with desirable properties. Automatic process controls and wafer transfer systems were implemented to alleviate the human operation generated control variation such as thickness, uniformity, and defects; and further to improve the systems' safety, reliability, and throughput. The requirements of semiconductor processing continuously evolve and drive the systems' performance improvements not only in CVD technology but also in productivity and cost, such as system level reliability, defects reduction, service ability, fabwide automation and

e-diagnostics, cycle time reduction from installation to production, environment friendly, and energy efficiency.

13.3.1 Summary of Widely Used CVD Reactor/Systems

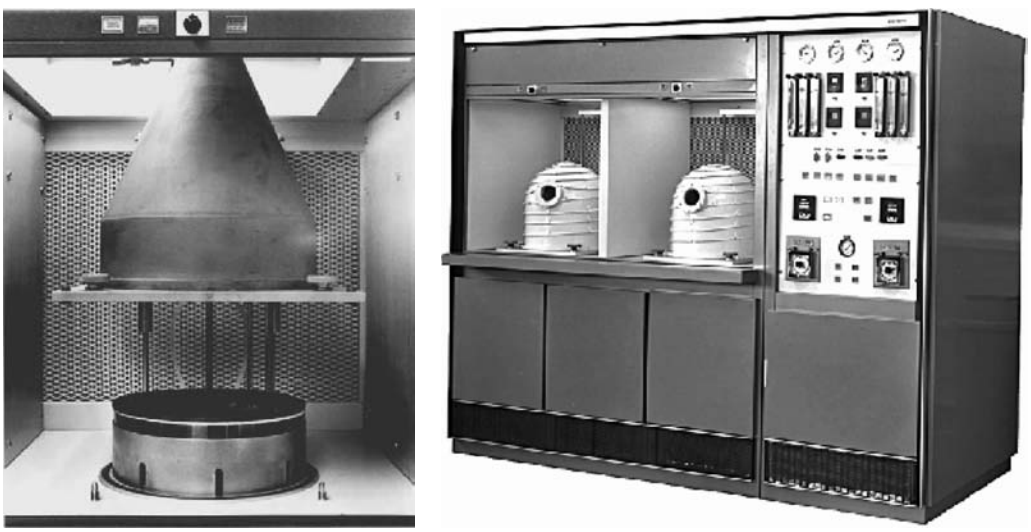
The following generally summarizes the types of CVD reactors in use since the 1960s

Hot wall systems
- Hot wall batch;
- Hot wall tube plasma reactor.

Cold wall systems
- Atmosphere continuous batch;
- Radiant heated atmosphere or reduced pressure batch;
- Low pressure batch;
- Batch plasma reactor;
- Multiple chamber, single wafer capacitive coupled plasma reactor;
- Multiple chamber, single wafer inductive coupled plasma reactor.

Most of the CVD reactors used in the 1960s and 1970s for integrated circuit (IC) fabrication were tube furnaces equipped with gas distribution inside the tube. Their advantages included excellent temperature uniformity close to ideal isothermal conditions (technology shared with thermal oxidation and annealing), particularly at temperatures $>600°C$, and high throughput with a batch load of wafers more than hundred, despite the low deposition rate. The drawbacks were that the hot wall surface received deposition as well as the wafers, eventually built-up thick deposits that would flake off and cause particles. Thus, the tube-type isothermal reactor was limited to high temperature ($>600°C$), low stress, good adhesion films such as polysilicon, high-temperature oxide TEOS or SiH_2Cl_2/N_2O, and high-temperature nitride (SiH_2Cl_2/NH_3). The lifetime of a tube furnace-type reactor was extended into 1980s by in situ cleaning capability such as the ClF_3 chemical clean to reduce the frequency of labor-intensive tube cleaning.

There are other types of CVD equipment which emerged during the late 1960s and early 1970s Bell-jar-type CVD systems, in some cases, they could accommodate multiple chambers (Figure 13.4). Continuous



(a) (b)

FIGURE 13.4 Early bell-jar-type CVD reactors (photo courtesy of Applied Materials).

belt conveyer-type atmospheric pressure reactors (Figure 13.5), wafer loading and unloading were easier than tube furnaces and the system architecture retained the high throughput advantage. Temperature control on the wafer was more challenging and this type of reactor was typically used for less temperature sensitive and more distribution sensitive chemistries such as LTO (SiH_4/O_2), and BPSG ($SiH_4/B_2H_6/PH_3/O_2$). The gas distribution system and the wafer carrier cleaning and maintenance were the weakest area for this type of system due to the chemistries used. A change in chemistry from SiH_4 base to $TEOS/O_3$ base reduced particle generation and extended the life of this type of system architecture, but the fundamental lack of in situ cleaning capability in the reactor limited its future.

Plasma deposition found its first application for silicon nitride passivation in the mid- to late-1970s (Figure 13.6). The superior diffusion barrier properties of nitride at low deposition temperature soon took over from LTO for the passivation layer. The parallel plate capacitively coupled plasma reactor is the simplest and easiest configuration. Easily applicable various frequencies of RF power, including low frequency, high frequency and dual frequencies can be used for fine tuning film properties. The versatile configuration and low temperature capability of parallel plate plasma reactor eventually become the dominant CVD reactor type today. In situ chamber cleaning capability also was introduced the first time

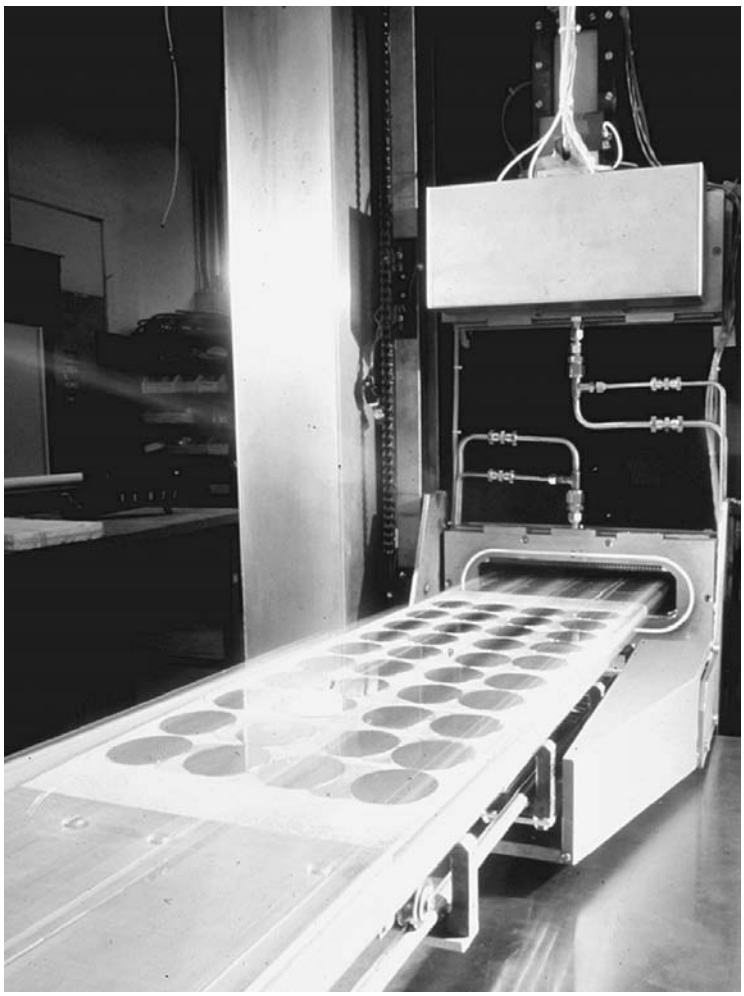


FIGURE 13.5 1970s conveyor-type atmospheric pressure CVD reactor (photo courtesy of Applied Materials).

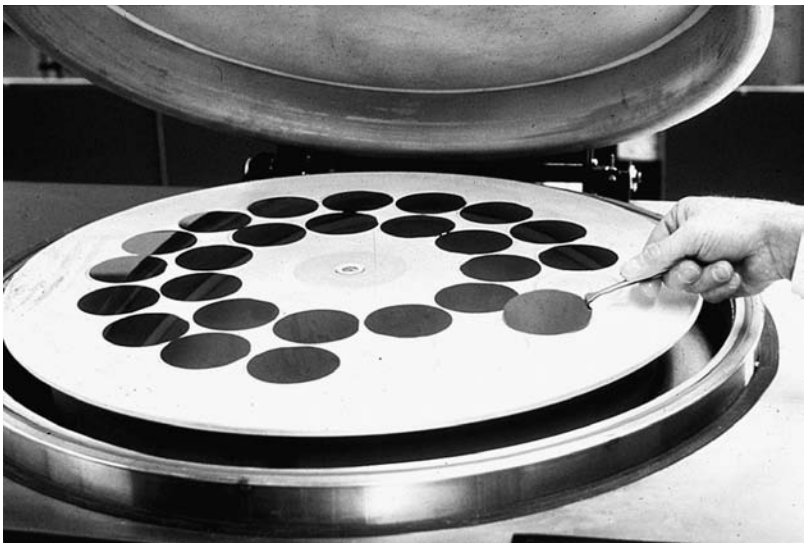


FIGURE 13.6 An early plasma CVD reactor, first introduced in 1976 (photo courtesy of Applied Materials).

onto the parallel plate plasma reactors. Wafer heating typically takes place through a resistive heater within the wafer pedestal. Radiation heating reactor as in tube furnace also can be converted to plasma reactor by inserting parallel electrode plates into tube furnaces, that extended the usable life of standard furnace technology, but the tube design still suffered from the same fundamental maintenance and cleaning issues and became obsolete for most of the applications.

In the mid-1980s, the need to reduce particles and the demand for lower-temperature deposition regimes that would be compatible with interlayer dielectrics for multiple level aluminum metallization initiated a new wave of reactors and systems designs. Computer process recipe control and automation of wafer transfer become available, resulting in more sophisticated process technologies. Plasma TEOS/O_2 for interlayer dielectric deposition, and plasma SiH_4 oxide and nitride deposition had extraordinary growth in their applications on multilevel interconnect.

Particle management, including a guarantee for particle performance, became essential. Plasma-enhanced chemical vapor deposition single-wafer, multiple-chamber systems (Figure 13.7) or multiple-station batch systems provided the required automation as well as in situ plasma cleaning capability to remove deposits on the heater or on chamber walls after each wafer or multiple wafers run. This capability significantly improved system uptime while maintaining the film properties and particle performance.

The same system architectures also enabled other applications and processes such as integrated deposition and etch in a single system, blanket tungsten, tungsten plug (integrated tungsten deposition and etchback), O_3/TEOS BPSG, tungsten silicide and polycide (integrated polysilicon tungsten silicide). Conventional batch systems in various configurations slowly lost their market share and limited in high temperature greater than 600°C applications. All of them are front end applications such as silicon epitaxial deposition, and gate nitride and oxide spacer deposition.

In the mid 1990s, the aspect ratio of the trenches between metal lines began to exceed the capability of plasma TEOS/O_2 even with deposition and etchback schemes. The demand for greater gap-filling capability resulted in the development of HDP deposition systems using either inductively coupled plasma or ECR plasma at low pressure ($<$10 mTorr), with bias power on the wafer. Ironically, HDP-CVD technology reverted back to the simple chemistry of SiH_4/O_2.

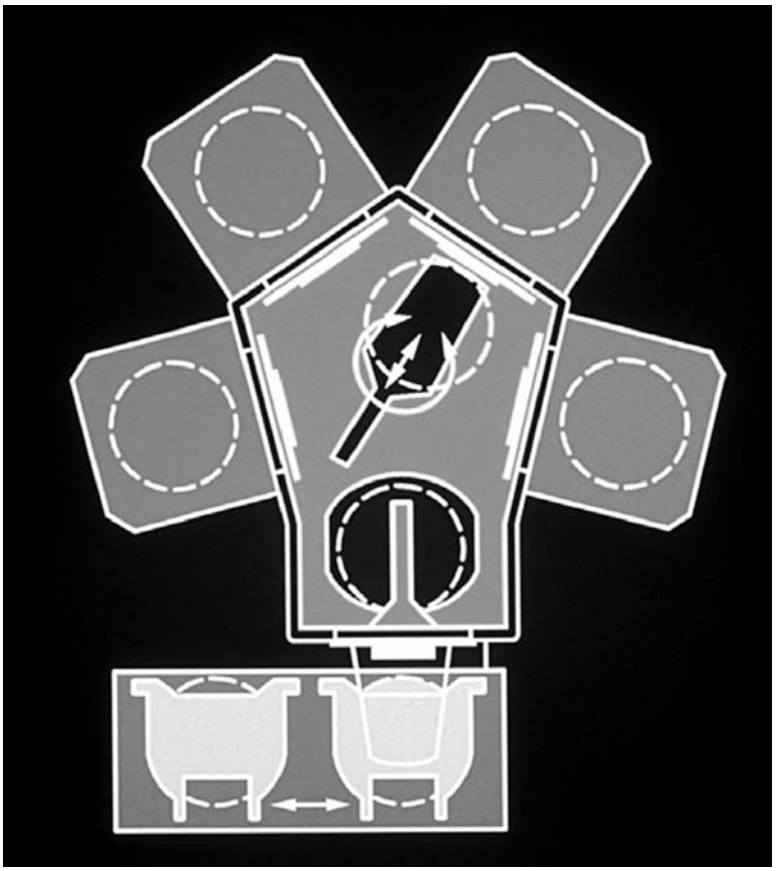


FIGURE 13.7 Schematic drawing of the single-wafer, multi-chamber precision 5000 CVD system (courtesy of applied materials).

At such as low pressure, wafer temperature control requires an electrostatic chuck to hold the wafer down and provide backside helium as the conduction medium for wafer cooling. High density plasma–CVD reactors still maintain in situ cleaning capability. These systems generally provide the capability of integrating several sequential depositions to form a complete dielectric layer. Figure 13.8 shows a high-density plasma reactor's process chamber.

High density plasma reactor did provide superior gap filling capability, but paid significant price of cost, in both capital and throughput. A new wave of technology change, copper dual Damascene with low k dielectrics, bounced the CVD reactor technology back to the old work horse: parallel plate plasma reactor.

The transition to 300 mm wafer size continued the trend of single wafer or small batch wafers processing. The scale up from 200 mm designs did not face any challenges. Temperature requirements, particularly the front end applications continuously drop. The thickness requirements also continuously drop. The variation of CVD materials is steadily increasing.

Future CVD reactor design will depend on the materials and their precursors chosen, as in the case of copper, low k interconnection and of emerging high k, metal gate transistor. The emergent class of reactor is for ALD, it requires a distinctive way of controlling gases introduced into the reactor and the delivery of each precursor for each application may have unique characteristics such as vapor pressure, stability, method of activation, and exhaust, while the basic reactor components remained conventional. The mechanical design, control, and location of the components handling the precursor will be added

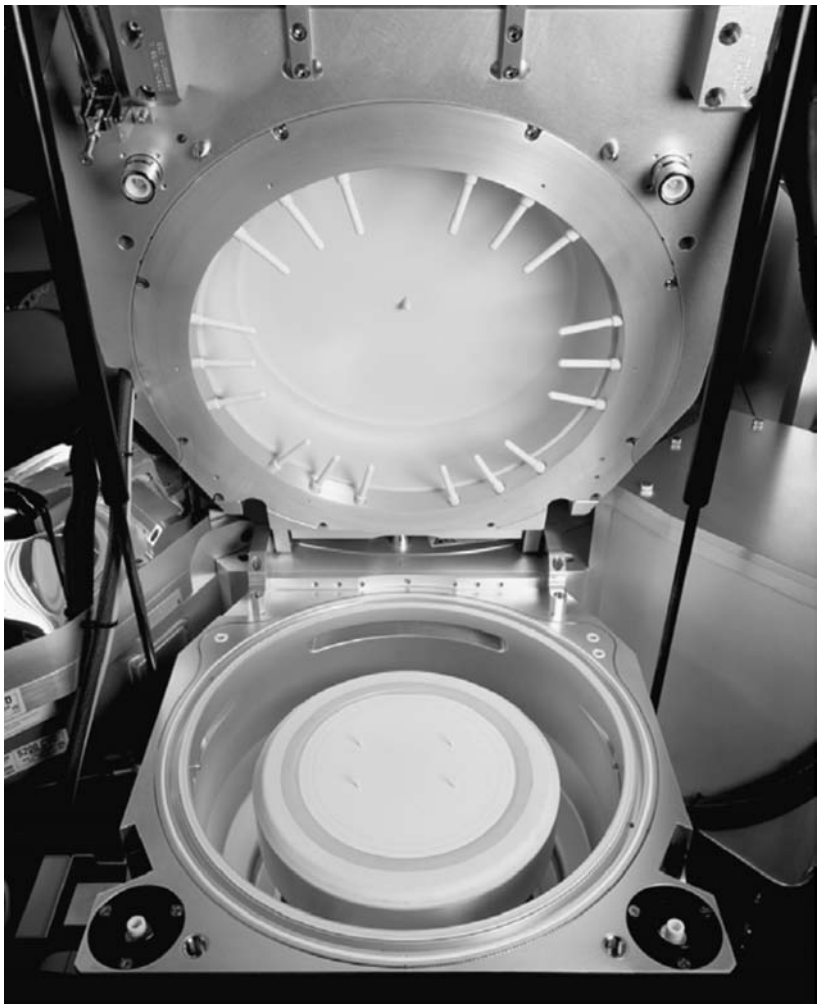


FIGURE 13.8 High density plasma (HDP)-CVD Single wafer processing chamber (photo courtesy of applied materials).

to the existing systems and/or reactors, while the existing technologies in wafer temperature control, gas distribution and chamber cleaning/particle control will be extended.

13.3.2 ALD Reactor Design: Next for CVD Equipment

Atomic layer deposition is a self-limiting process. The process is much less sensitive to gas flow and temperature uniformities. Therefore, the chamber hardware requirements for running ALD processes are not as critical as CVD. In many cases, initial ALD chemistry screening and process characterization can be performed using existing CVD reactors without major hardware modifications. Once process chemistry is determined, the chamber hardware designs need to be optimized based on process specifications to obtain the best performance.

A common requirement for semiconductor manufacturing equipments is cost effectiveness. In other words, the chamber must be low on the cost of consumable and high in productivity. To meet such requirements, high precursor utilization rates and low cycle time must be achieved. To maximize

precursor utilization efficiency and to minimize pulsing time, industrial designed ALD reactors are normally operated at pressures about 1 Torr or above to increase reaction probability. A laminar inert gas flow is applied to carry precursors through the reactor and to purge un-reacted precursors and reaction byproducts. The chamber should be designed with small volume as possible and minimize flow restrictions such that each process step can be completed rapidly to reduce cycle time.

In ALD, the separation of different precursor exposures reduced the risk of particle generation by gas phase reaction. The highly reactive nature of ALD precursors and the requirements of exposure saturation greatly lowered the deposition temperature. As a result, chamber inner walls can be coated even in a cold wall reactor. The use of moving parts inside ALD reactor is highly undesired. In addition, many of ALD processes use liquid or solid precursors with low vapor pressure as source gases. The source chemicals need to be heated to temperatures where appropriate vapor pressures can be obtained. Special heating of gas line and valve designs will be needed to maintain proper temperature to maximize the efficiency of precursor vapor delivery. A cold trap is required to prevent precursor condensed in exhaust pipes and vacuum pumps causing particle and vacuum pump reliability issues.

Both batch and single reactor designs can be applied for ALD processes. The adoption of chamber designs depends on the process flow in device manufacturing steps. For example, batch designs can be used for front end high *k* gate dielectric applications where vacuum break is tolerable. In contrast, most of metal processing including metal gate requires single chamber designs to fit different types of process chamber onto a cluster tool for multiple layer films stack depositions without breaking vacuum to maintain interface integrity.

In summary, the main parts on designing ALD chambers are:

- No strict precursor flow uniformity requirement;
- No highly uniform substrate heating is required;
- Exposure and purge sequences need to be completed rapidly;
- Precursor utilization efficiency;
- Avoid using moving parts inside ALD reactors;
- Gas line heating uniformity if liquid or solid precursors are used;
- Exhaust pipe and vacuum pump maintenance.

13.4 CVD Thin Films

13.4.1 Dielectrics

13.4.1.1 Silicon Dioxide

Silicon dioxide (SiO_2) has dominated semiconductor processing for several decades as the principle insulator between polysilicon and multi-level metal interconnects final passivation layers. Its unique properties include: (i) high mechanical strength; (ii) good adhesion to the underlayers; (iii) high electrical resistance and high breakdown voltage; (iv) impermeability to moisture and alkali metals; and (v) high chemical, radiation, photoactive, and thermal stability.

SiO_2 films are normally formed by either high-temperature furnace oxidation or by CVD. Unlike thermal oxidation, CVD normally requires a silicon-containing precursor and produces an amorphous SiO_4 tetrahedral structure [1], with an empirical formula SiO_2, and is therefore called silicon glass [2,3].

With more process controlling variables, CVD is capable of producing SiO_2 films with various chemical, mechanical, and electrical properties, including film density, etch rate, step coverage, substrate sensitivity, stress and dielectric strength, etc. Relatively, high deposition temperature as well as post-deposition annealing can modify the properties of CVD films towards those of thermal oxides. Thermal CVD has a silicon source, e.g., TEOS or SiH_4, whereas thermal oxidation is simply oxidizing silicon surfaces. Therefore, only O_2 or steam is used for thermal oxide growth. Since the substrate silicon is in crystal configuration, the film properties are very stable. There is no gas phase nucleation for thermal oxide growth. Furthermore, dopants can be added into SiO_2 films during CVD processing, in order to

TABLE 13.2 CVD Reactors for Silicon Oxide Deposition

Reactor	PECVD	APCVD	SACVD	LPCVD
Pressure (Torr)	1–10	760	50–700	<10
Temperature (°C)	200–550	<500	<600	300–900
Reaction type	RF plasma	Thermal	Thermal	Thermal
Reactor type	Single wafer	Continuous belt	Single Wafer	Furnace

form binary and ternary silicates with unique properties. Therefore, besides its common application in semiconductor manufacturing as insulation, silicon glass is also used as getters [4], as diffusion sources [5–7], as diffusion, implantation and etch masks, as diffusion barrier layers [8], and as sacrificial material [9].

Based on the choice of reaction and process conditions (temperature, pressure, reactant flow, system configuration, etc.), various types of CVD equipment have been developed to deposit SiO_2 films, as listed in Table 13.2.

13.4.1.1.1 Plasma-Enhanced CVD (PECVD) and High Density Plasma CVD (HDP-CVD)

For PECVD processing, the use of energy in a plasma state creates more reactive radicals, allowing reactions to occur at lower temperatures. It also results in high film density and compressive stress due to high-energy ion bombardment. Higher deposition temperature is often desired to further enhance film density and drive off volatile species. In principle, the reactions take place only within the plasma region, which therefore reduces residue formation on reactor walls. However, if the deposition process or reactor hardware is not optimized, charged species and intermediate products may be trapped inside the film during deposition, leading to high impurity levels and residual charging problems. Furthermore, excess ion bombardment may cause plasma damage to the device as well as to reactor hardware. Step coverage for the PECVD process is controlled by adding a bias voltage to the substrate to attract the charged radicals into the bottom of the gap. Therefore the gap-fill performance strongly depends on deposition conditions and hardware configuration.

Even with the addition of DC bias, the step coverage for conventional PE oxide films is still less than that of thermal CVD oxides (Figure 13.9a). As device dimension keeps shrinking, this becomes more and more important. As an alternative approach to use the existing PECVD tool, the ex situ Dep/Etch/Dep process was developed to overcome the poor step coverage while maintaining the advantages from PE films. The concept is illustrated in Figure 13.7, whereas Ar ions are utilized to sputter off the excess film on top of the gap; a 45° surface angle has the maximum sputter yield (Figure 13.9b). Besides

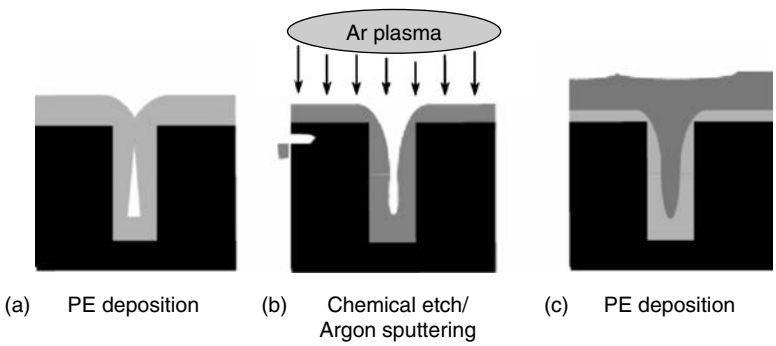


FIGURE 13.9 Schematics for dep/etch/dep processes.

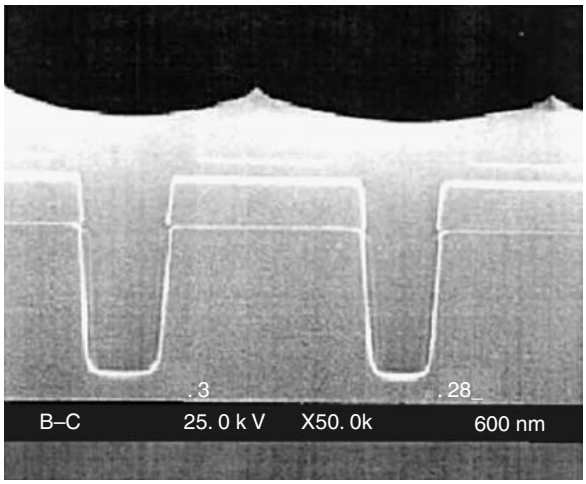


FIGURE 13.10 HDP gap fill performance for shallow trench isolation application.

re-deposition to the bottom of the gap to help in subsequent gap filling, the sputter action opens up the voids for subsequent PE deposition to fill the gap completely (Figure 13.9c).

The amount of etch depends on gap geometry and multiple dep/sputter steps may be required in order to prevent corner clipping. The problems associated with multi-process steps and huge throughput loss led to the development of HDP-CVD reactor, where film deposition and Ar sputtering can be carried out simultaneously in a single process tool.

Both reactant, e.g., SiH_4 and O_2, and Ar gases were introduced into the process chamber. While SiO_2 film is deposited through plasma enhanced reaction, Ar ions are also formed inside the plasma and directed to the substrate via bias RF power. The resulting high energy ion bombardment causes physical sputtering of the deposited film. Since the degree of sputtering is controlled by both ion density and ion energy, high sputter rate can be achieved using HDP. The deposition to sputter ratio can be adjusted by varying the process controlling parameters to achieve the best gap fill performance, however due to the same fashion, high aspect gap fill may result in lower throughput. Superior gap fill performance, high film density and minimal metal contamination have qualified HDP oxide for shallow trench isolation (STI) applications without any post deposition anneal. Besides, its application also involves intermetal dielectric due to low deposition temperature, pre-metal dielectric (PMD) while doped with phosphorus and passivation oxide. Displayed in Figure 13.10 is the HDP gap fill performance for STI application. The triangular shape on top of each gap is a characteristic signature of the deposition/sputter process. Normally SiH_4 chemistry is used in HDP-CVD reactor to yield minimal impurity level, because the gap fill performance is controlled by physical sputtering instead of deposition chemistry.

13.4.1.1.2 Low Temperature Oxide

Unlike PECVD or thermal SiO_2 films, which have compressive stress, low-temperature CVD oxide films are normally deposited with tensile stress, ranging from 1×10^8 to 3×10^9 dynes/cm^2, depending on process conditions. Moreover, they also exhibit lower film densities, high etch rates in buffered hydrofluoric acid (BHF) and have a refractive index of 1.44. Table 13.3 shows a comparison of the film properties among these oxides:

Subsequent high-temperature annealing at 700°C–1000°C causes densification of the oxide films, and leads to chemical properties similar to thermal oxides. Such a densification process not only changes film thickness and film density from 2.1 to 2.2 g/cm^3, but also causes chemical reconstruction of the film [20], leading to compressive film stress and low HF etch rate. High-temperature annealing also removes H

TABLE 13.3 Properties of CVD and Thermal Silicon Dioxides

	Thermal Oxide	PECVD Oxide	LTO Oxide
Etch Rate (6:1 BHF, Å/min)	900	1200	2700–9000
Shrinkage (1000°C)	0%	2%	4%–10%
Film Stress ($\times 10^9$dynes/cm^2)	2–3 (C)	0.1–3 (C)	0.1–3 (T)
Refractive index	1.462	1.46–1.48	1.444

contamination which is incorporated into the film both during deposition, including PECVD, and post-deposition through moisture absorption. Before annealing, these hydrogen atoms are bonded to the SiO_2 network as Si–H, Si–OH, and H–OH, with a concentration of 2%–20% [19–21] depending on film type and process conditions.

13.4.1.1.3 Low-Pressure CVD (LPCVD)

Another method to achieve SiO_2 film with properties close to thermal oxide is to deposit the film at high temperature in LPCVD reactors. One example is the reaction between dichlorosilane and nitrous oxide at nearly 900°C [22]:

$$SiH_2Cl_2(g) + 2N_2O(g) \xrightarrow{T=900°C} SiO_2(s) + 2N_2(g) + 2HCl(g) \tag{13.1}$$

This process yields a high-quality oxide with good step coverage. However, it suffers from low deposition rates and can be used only on polysilicon surfaces before Al deposition due to high temperature. In addition, chlorine contamination is also a concern [16].

13.4.1.1.4 Dielectric CVD Chemistries

13.4.1.1.4.1 Silane (SiH_4)

Before the introduction of TEOS [10,11], the silane-based reaction was widely used to form CVD SiO_2 films in semiconductor manufacturing. The so-called LTO utilizes the chemical reaction between silane and oxygen in the temperature regime between 300 and 450°C. The reaction is given by:

$$SiH_4(g) + O_2(g) \xrightarrow{T=300°C–450°C} SiO_2(s) + 2H_2(g) \tag{13.2}$$

Dopants can be incorporated into the SiO_2 film, with the addition of PH_3 and/or B_2H_6 parallel reactions, to form PSG or BPSG:

$$4PH_3(g) + 5O_2(g) \xrightarrow{T=300°C–450°C} 2P_2O_5(s) + 6H_2(g) \tag{13.3}$$

$$2B_2H_6(g) + 3O_2(g) \xrightarrow{T=300°C–450°C} 2B_2O_3(s) + 6H_2(g) \tag{13.4}$$

Due to the high reactivity between silane and oxygen, a low-temperature regime (300°C–450°C) is usually chosen in order to control the reaction rate to achieve good film thickness uniformity. The activation energy is less than 9.2 kcal/mol depending on O_2:SiH_4 ratio, which indicates a gas phase diffusion or surface adsorption controlled process. The conventional CVD tools for these reactions include atmospheric pressure chemical vapor deposition (APCVD), LPCVD, and PECVD reactors.

For thermal CVD reactions, the SiO_2 is formed through a gas surface heterogeneous reaction. However, homogeneous gas phase nucleation happens simultaneously, forming SiO_2 white powder in the reaction chamber and potentially creating particle contamination in the deposited film. Therefore, either a cold wall reactor or a lower pressure regime is preferred to reduce particle contamination.

On the other hand, silane can also react with O_2 or N_2O in a plasma environment at temperature less than 400°C.

$$SiH_4(g) + 2N_2O(g) \xrightarrow{T=200°C–400°C,\ RF\ \ Plasma} SiO_2(s) + 2H_2(g) + 2N_2(g) \tag{13.5}$$

Impurities, such as hydrogen or nitrogen, are often incorporated into the films during the PECVD process, which may alter the oxide properties. The refractive index of the oxide film can vary due to the impurity level, as well as the Si:O ratio. Nearly stoichiometric ($n = 1.46$) oxide films can be achieved using the SiH_4/O_2 plasma reaction.

13.4.1.1.4.2 TEOS

Even though silane chemistry is simple, the resulting film suffers from poor step coverage [12,13], especially for sub-micron gap spaces. Therefore, silane has been largely replaced by TEOS during the last decade in order to deposit oxide films with good conformality in those applications requiring it. At temperatures above 600°C, TEOS decomposes and forms SiO_2 film [14,15], with or without oxygen:

$$Si(OC_2H_5)_4(vapor) \xrightarrow{T>600°C} SiO_2(s) + H_2O(g) + 2C_2H_4(g) \tag{13.6}$$

The reaction pyrolysis does not require additional O_2 to yield stoichiometric SiO_2 film, since the TEOS molecule contains oxygen. A LPCVD tool is the best candidate for such a reaction due to its high temperature capability. Tetraethylorthosilicate films exhibit an optimum deposition rate and film thickness uniformity at various temperatures [16], which yields the optimum deposition temperature regime (680°C–730°C) with an apparent activation energy of 40 kcal/mol. Above 730°C, TEOS depletion begins, due to the shift in the process controlling step.

13.4.1.1.4.3 TEOS/O_3

With the addition of ozone [17] or using the plasma-enhanced process [11], the deposition temperature for TEOS decomposition can be dramatically reduced. The overall reaction can be expressed as follows, which can be carried out in PECVD, APCVD as well as SACVD tools:

$$Si(OC_2H_5)_4(vapor) + 8O_3(g) \xrightarrow{T=400°C–600°C} SiO_2(s) + 10H_2O(g) + 8CO_2(g) \tag{13.7}$$

$$Si(OC_2H_5)_4(vapor) + 12O_2(g) \xrightarrow{T=400°C,\ RF\ \ Plasma} SiO_2(s) + 10H_2O(g) + 8CO_2(g) \tag{13.8}$$

However, the actual reaction mechanism is much more complicated and is still under investigation. For the TEOS/O_3 reaction, it is generally accepted that TEOS decomposes to some kind of intermediate in the gas phase, which then adsorbs onto the substrate and further decomposes to produce the SiO_2 film [18]. Due to high surface mobility of the intermediate, the resulting oxide film has excellent conformality.

Dopants can be incorporated into the TEOS/O_3 SiO_2 film with the addition of organo-dopant precursors into the reaction mixture, e.g., TEPO, trimethylphosphate, TEB, and trimethylborate, replacing the traditional hazardous and non-stable hydrides (B_2H_6 and PH_3). Unlike the silane reaction, the addition of dopants changes the reaction mechanism and thus alters the film properties. Displayed in Figure 13.11 are the temperature and pressure effects for both undoped and doped TEOS/O_3 films using TEB and TEPO [19].

The undoped silicate glass (USG) process is characterized by a strong temperature effect with an apparent activation energy of −24.1 kcal/mol and weak pressure dependence, indicating the dominance of the surface reaction; whereas BPSG deposition is just the opposite, indicating a gas phase diffusion limited process, which results in worse step coverage.

13.4.1.1.5 Applications and Properties of Undoped Silicon Glass

Even though undoped CVD SiO_2 has many similar properties to thermally grown oxides, it also has its unique features, which diversifies the applications which it can be used for. Excellent step coverage and high deposition rate at low temperature qualify its use as an inter-metal dielectric (IMD) for multi-level IC structures. Since aluminum cannot tolerate temperatures higher than 450°C, neither a high-temperature LPCVD reaction nor thermal oxide can achieve this task due to their extremely low

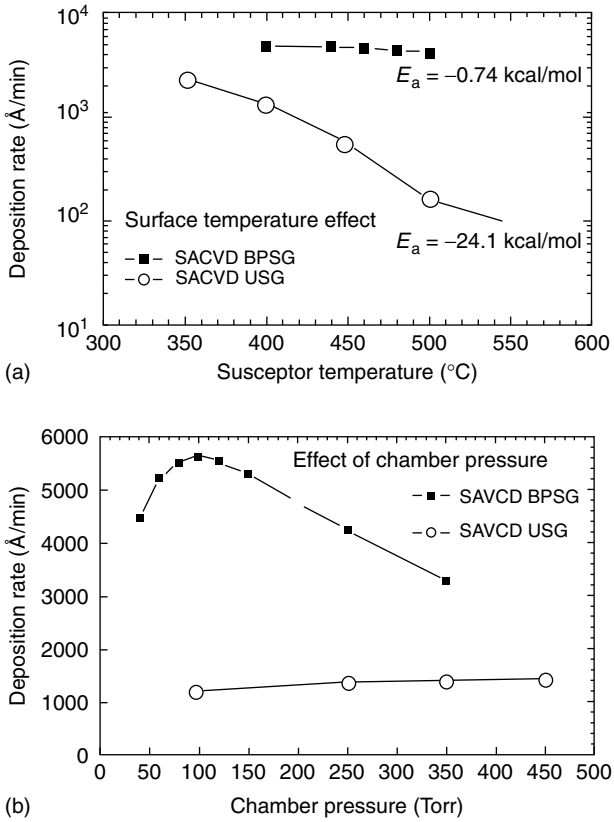


FIGURE 13.11 The effect of (a) substrate temperature and (b) chamber pressure on the deposition rate of the SACVD undoped silicate glass and borophosphosilicate glass (BPSG) processes.

deposition rates. Even for PMD, lower thermal budget is required to reduce dopant diffusion as device dimensions keeps shrinking.

Gap filling for CVD USG film is mainly achieved through conformal film growth, as illustrated in Figure 13.12. Conformality is usually the measure for deposition rate uniformity over all substrate topography (Figure 13.12a). In gap-filling applications, as the deposition process goes on, the film merges at the center from both sides as well as at the bottom and thus fills the gap (Figure 13.12b).

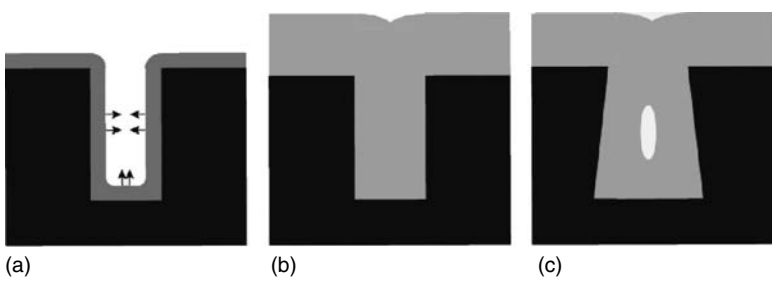


FIGURE 13.12 Schematic for conformal gap fill.

TABLE 13.4 Effects of Process-Controlling Variables on SACVD USG Film Properties (↑ increase; ↓ decrease)

Effects	Deposition Rate	Etch Rate	Shrinkage	Step Coverage
Chamber pressure ↑	No Effect	↓	↓	↑
TEOS flow↑	↑	↑↑	↑↑	↓
O_3 flow↑	↓↓	↓↓	↓↓	↑↑
O_3 concentration ↑	↓↓	↓↓	↓↓	↑↑
Substrate temperature ↑	↓	↓↓	↓↓	↑

Tetraethylorthosilicate-based CVD film can reach close to 100% conformality [2,16,30,31] due to the fact that the process is controlled by surface reaction and is independent of reactant supply from the gas phase [18,19]. Therefore, the gap-filling performance is independent of the aspect ratio of the gap. On the other hand, as the sidewall profile becomes vertical or negative (Figure 13.12c), the film tends to close at the top, preventing further deposition inside the gap and causing voiding problems.

As mentioned before, film properties of the CVD oxide can be tuned by process conditions; Table 13.4 shows an example using the SA CVD ozone/TEOS process [20]. Likewise, PECVD processes can also yield various film properties.

Therefore, CVD films can be optimized for different applications, including:

- High aspect ratio gap fill due to the exceptional film conformality and global planarization capability;
- PE USG liner as a moisture barrier [8];
- Sacrificial films with high etch rate;
- Sidewall spacer due to excellent step coverage;
- Hard mask with critical dimension (CD) control;
- Deep UV (DUV) and dielectric anti-reflective coating (DARC) due to minimal reflectance and superior film uniformity;
- Film stack with nitride due to tunable film stress;
- Shallow trench isolation [20,23,24] due to gap filling capability, low metal contamination and chemical–mechanical polishing (CMP) compatibility, replacing traditional local oxidation of silicon (LOCOS) to prevent "bird's beak" [25,26]. (Normally, the STI process requires post-deposition annealing to enhance the film etch resistance [20]).

Despite all the advantages, CVD oxides have their own drawbacks; low film density and high etch rate for low-temperature films, and high impurity levels (H, C, etc.) for plasma-enhanced processes. Therefore, high-temperature processes still dominate in the areas of gate oxides and field oxides.

Furthermore, ozone/TEOS processes exhibit surface sensitivity, i.e., change in film's properties as a function of substrate structure [27,28]. This is mainly due to the fact that the deposition is controlled by a surface reaction process [19], and the sensitivity can be reduced by lowering the Ozone/TEOS ratio or pressure. Unfortunately, these changes in process conditions usually lead to worse gap-filling performance. On the other hand, however, this unique feature also enables so-called selective oxide growth [29] to achieve self-planarization during the STI process due to the difference in deposition rate and film density on nitride and silicon substrates.

13.4.1.1.6 Applications and Properties of Borophosphosilicate Glass

Borophosphosilicate glass is a $B_2O_3/P_2O_5/SiO_2$ ternary silicate glass, which can be formed with the addition of boron and phosphorus dopants during deposition. The major benefit of dopant incorporation is the change in surface tension of the film. As a result, BPSG film becomes viscous at elevated temperatures and responds to surface tension forces, rounding sharp corners, which smoothens the surface topography and achieves void-free gap fill. The reflow process can take place either in a furnace or via rapid thermal annealing at temperatures >700°C, depending on the dopant concentration, film density, and annealing environment (temperature, time, and ambient) [32–36].

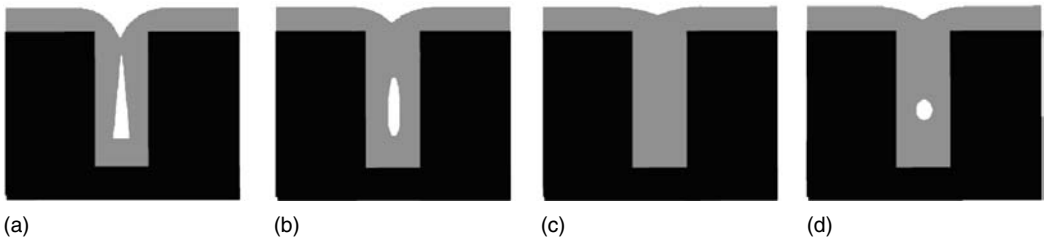


FIGURE 13.13 Schematics for BPSG reflow gap fill.

Lower annealing temperatures can be achieved with higher dopant concentration under steam ambient. However, increase in P concentration above 5-wt.% does not further enhance BPSG reflow capability, whereas excess B causes the film to be very hygroscopic and unstable. After anneal, BPSG films with total dopant concentration above 11-wt.% or high B concentration ($>$6-wt.%) tend to form BPO_4 crystals on the surface [37].

The gap-filling mechanism for BPSG film is illustrated in Figure 13.13. Due to change in the reaction-controlling step to gas phase diffusion [19], the as-deposited BPSG film is less conformal, compared to a USG film, and can leave long voids inside the gap (Figure 13.13a). During anneal, the film reflows and voids start to shrink, as shown in Figure 13.13b. If reflow is sufficient, the gap will be filled completely (Figure 13.13c), otherwise, circular voids or pin holes will result (Figure 13.13d).

With the optimization of the deposition process, as well as post-deposition anneal, minimal thermal budget is required to fill high aspect ratio gaps, even with negative sidewall profiles. Displayed in Figure 13.14 is the gap fill performance using an SACVD 2-step BPSG process to achieve void-free gap fill [38] at 0.06 μm spacing and 6:1 aspect ratio.

Since BPSG gap-fill normally requires a post-deposition anneal, it is mainly used as a PMD due to temperature constraints. Like all CVD processes, film properties can change depending on deposition conditions. In addition to its reflow capability, dopants also change the chemical properties of the film. Displayed in Figure 13.15 is the wet etch rate for BPSG film in different etch solutions. Etch rate for

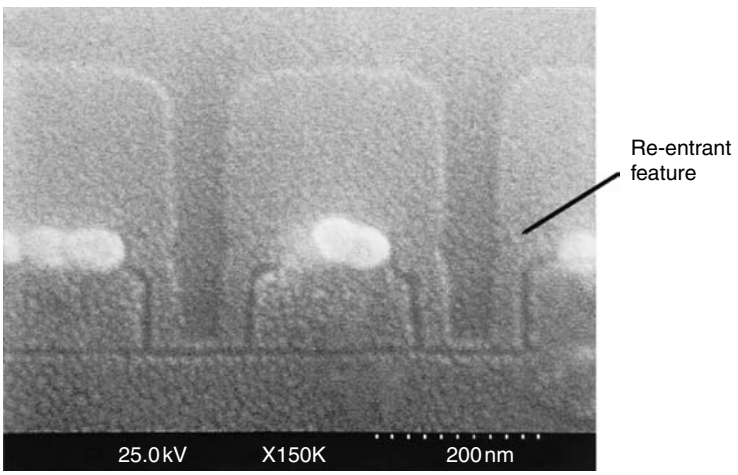


FIGURE 13.14 Gap fill performance using an SACVD 2-step BPSG process to achieve void-free gap fill at 0.06 μm spacing and 6:1 aspect ratio.

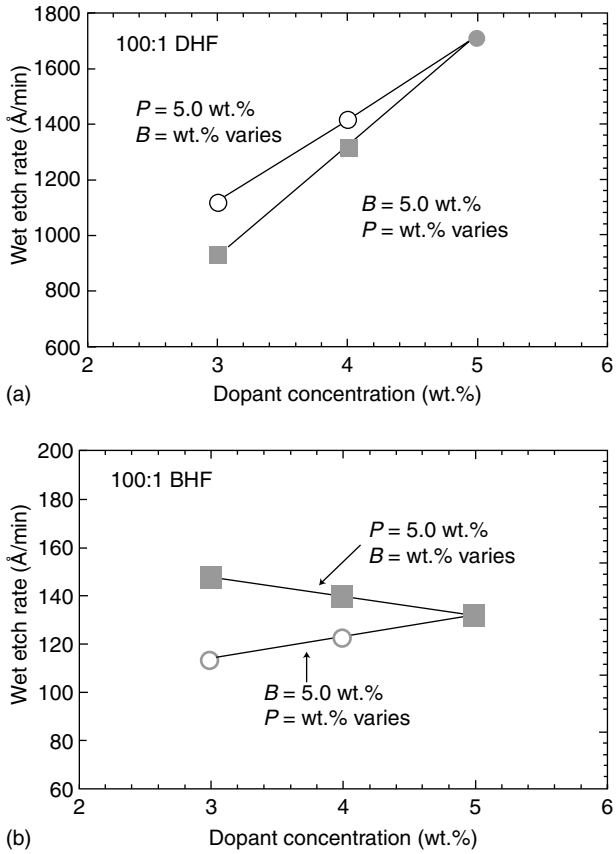


FIGURE 13.15 SACVD BPSG wet etch rate in (a) 100:1 diluted HF and (b) 100:1 buffered HF.

higher B concentration films increases in diluted HF and decreases in buffered HF solution. Similar to the USG film, BPSG etch rate decreases after anneal due to film densification; stress temperature hysteresis also exhibits the same behavior [20,36].

Due to the incorporation of phosphorus, BPSG film can also be used to getter alkali ions, similar to the behavior of PSG [39,40]. However, the major problem for BPSG is dopant diffusion into the silicon underlayer, especially phosphorus outdiffusion, which increases with higher boron concentration.

13.4.1.1.7 Applications and Properties of Boron (BSG) or Phosphorus (PSG) Doped Silicon Glass

Both BSG and PSG belong to the binary silicate glass category and have their unique features. With their tendency for dopant outdiffusion, both films can serve as a diffusion source [5,41]. This is extremely useful in forming uniform doping inside gaps with high aspect ratio, because the deposition chemistry is capable of forming conformal oxide films. Ultra-shallow doping profiles can be achieved via rapid thermal processing (RTP) to drive in dopants [6,42]. For the same reason, however, a dopant diffusion barrier is normally required if any doped silicate glass is utilized as insulator between polysilicon.

Another major characteristic of the PSG film is its alkali ion gettering capability [39,40], which makes it well suited for use as a passivation layer. With shrinking device geometries, the conformality of the PSG layer becomes more critical and the film is often needed to fill the gaps directly. Unfortunately, the step coverage for PSG film is poor compared to USG film due to the similarities with BPSG in the reaction-controlling step, even with ozone/TEOS chemistry. In order to achieve void-free gap fill, the gap sidewall

is usually slightly tapered to compensate for the diminished conformality. Increasing the deposition temperature as well as lowering the P concentration enhances PSG gap-fill performance. Post-deposition annealing above 1000°C at pressures from 1 to 25 atm causes PSG film to flow [43] due to change in viscosity; however, this is rarely used due to device thermal budget constraints.

Similar to BPSG film, the etch rates for BSG and PSG vary depending on dopant level and etch solutions, as well as deposition conditions. They can be optimized to yield high wet etch selectivity to thermal oxide [44], while maintaining dry etch selectivity to silicon. Therefore they can be used as hard masks for silicon deep trench etch or as sacrificial layers [9]. Unlike BPSG film, the as-deposited BSG and PSG films are porous and become increasingly hygroscopic at high dopant levels. Normally, dopant concentrations higher than 6.0 B-wt.% or 8.0 P-wt.% will induce boric or phosphoric acid formation and lead to film instability. As-deposited film density can be enhanced by using a PECVD process or high-deposition temperature, as well as post-deposition densification [45].

13.4.1.2 Silicon Nitride

Silicon nitride (Si_3N_4) is another dielectric material widely used in very large-scale integration VLSI fabrication. Its importance to modern technologies lies in its impermeability to most impurities, which qualifies its primary use as a passivation layer, especially as a diffusion barrier to moisture and sodium, or as a selective oxidation mask to prevent oxygen from penetration into the silicon underlayer. Due to its high dielectric constant (6–9 vs. 4.0 of CVD silicon oxide), it is less attractive as an insulator, because the resultant higher capacitance between polysilicon and metal interconnects leads to a larger RC delay. On the other hand, this also makes it a better candidate for the capacitor dielectric as well as gate dielectric to reduce device geometry [46,47]. Furthermore, the difference in molecular structure and composition leads to the chemical and physical properties of Si_3N_4 films, compared to SiO_2, in term of etch rate (10–15 Å/min for LP nitride vs. 900 Å/min for steam oxide in 6:1 BHF), film density (2.5–3.1 vs. 2.2) as well as thermal expansion (4×10^{-6}/K vs. 5.6×10^{-7}/K). Therefore, silicon nitride is also used as an etch hard mask or CMP stop in silicon deep trench etch [48,49] or STI applications [24,25] to remove or planarize silicon oxide. Because of the thermal mismatch between silicon and Si_3N_4, silicon nitride is usually deposited onto Si with a thin oxide buffer layer, called pad oxide, to prevent stress-induced damage at the interface at elevated temperature [24,25].

Silicon nitride is an amorphous material, similar to SiO_2 films, and can be formed mainly through deposition rather than furnace growth, including LPCVD and PECVD or recently HDP-CVD. Furnace-grown Si_3N_4 is achievable at temperatures above 1000°C on silicon with N_2 or NH_3. However, the hermeticity of the growing nitride film to the reactant species prevents further nitridation and limits its thickness to less than 50 Å. APCVD is also possible through the following reaction to achieve stoichiometric composition:

$$3SiH_4(g) + 4NH_3(g) \xrightarrow{700°C–900°C} Si_3N_4(s) + 12H_2(g) \tag{13.9}$$

However, the reaction rate is less than 10 Å/min at 700°C and increases to 1000–2000 Å/min at 900°C [50,51]. Less controllable film thickness uniformity due to strong temperature dependence, as well as high particulate contamination due to excess gas phase nucleation, limit the practical value of this process. Low-pressure chemical vapor deposition becomes a more attractive deposition method, which utilizes the following reaction chemistry with a deposition rate of 15–20 Å/min:

$$3SiH_2Cl_2(g) + 4NH_3(g) \xrightarrow{700°C–800°C} Si_3N_4(s) + 6HCl(g) + 6H_2(g) \tag{13.10}$$

The film properties are listed in Table 13.5, compared with PECVD Si_3N_4 film [9].

As can be seen, LPCVD nitride yields high quality nitride with stoichiometric Si:N ratio, minimal H impurity, low etch rate and excellent step coverage. However, thick films (>2000 Å) cannot be deposited due to not only low deposition rate, but also highly tensile stress causing film cracking. High deposition

TABLE 13.5 LPCVD and PECVD Silicon Nitride Film Properties

Film Properties	LPCVD	PECVD
Composition	Si_3N_4	$Si_xN_yH_z$
Si:N ratio	0.75	0.8–1.0
Dielectric strength (V/cm)	1×10^7	6×10^6
Film density (g/cm^3)	2.8–3.1	2.5–2.8
Refractive index	2.0–2.1	2.0–2.1
Dielectric constant	6–7	6–9
Bulk resistivity (W/cm)	10^{15}—10^{17}	10^{15}
Surface resistivity (W/Sqr)	$>10^{13}$	10^{13}
Stress on Si (Dyne/cm^2)	$1.2–1.8\times10^{10}$ (Tensile)	$1–8\times10^9$ (Compressive)
Thermal expansion (K^{-1})	4×10^{-6}	$4–7\times10^{-6}$
Step coverage	Conformal	Conformal
H_2O permeability	Zero	Low/none
Thermal stability	Excellent	Variable >400°C
Na^+ penetration	<100 Å	<100 Å
Na^+ retained in top 100 Å	>99%	>99%
IR absorption (cm^{-1})	Si–N (870)	Si–N (830), Si–H (2180)
Wet etch rate (Å/min)		
6:1 BHF (20°C–25°C)	10–15	200–350
49% HF (23°C)	80	1500–3000
85% H_3PO_4 (155°C)	15	100–200
85% H_3PO_4 (180°C)	120	600–1000
Plasma etch rate (Å/min)		
70% CF_4/30% O_2, 150 W, 100°C	200	500

temperature also limits the use of LPCVD Si_3N_4 to films deposited prior to any metal deposition. Therefore, applications of LP nitride mainly consist of capacitor dielectric with low leakage current and high breakdown voltage [46,47]; a composite gate dielectric while stacking with thermal oxide [52,53]; a spacer to protect oxide passivation edges for moisture sealing [54]; a hard mask for LOCOS [55] or deep silicon trench etch [48,49] and STI oxide CMP stop [24] with high etch selectivity to SiO_2 films; and as a masking layer for selective oxidation [56]. Due to the nature of the reactor type, LP nitride produces film deposition on both front and backside of the wafers. Therefore, unlike PE nitride, additional etch steps are normally required after LP nitride to remove the backside film.

In a PECVD tool, the reaction temperature to form Si_3N_4 film can be dramatically reduced. Since PECVD could potentially trap reaction intermediates inside the film, SiH_4 chemistry is adopted to achieve high deposition rate (100–300 Å/min) and low impurity level. The overall reaction is as follows:

$$SiH_4(g) + NH_3(g) \quad \text{or} \quad N_2(g) \xrightarrow{T=200°C–550°C,\ \text{RF Plasma}} Si_xN_yH_z(s) + H_2(g) \tag{13.11}$$

The film properties for PE nitride can vary with different process conditions, i.e., pressure, NH_3 to SiH_4 ratio, wafer temperature, RF power, etc. Good step coverage can be achieved with the external DC bias or in situ sputter-deposition for the HDP-CVD process. High-energy plasma bombardment causes film stress to be compressive and tunable at different RF power levels to minimize film cracking. However, as shown in Table 13.1, silicon-rich films usually result from PECVD nitride deposition, which leads to lower film density and high etch rate. H contamination is also a major concern, which is a strong function of deposition temperature. Nuclear reaction analysis reveals that the Si_3N_4 film deposited at 400°C has 25 at.% of H, decreasing to 16 at.% at 480°C, and 13 at.% at 550°C. High H content can cause significant threshold voltage shift in IC devices as well as affecting etching characteristics.

PE nitride is mainly used as the diffusion barrier layer or final passivation layer. When the device thermal budget is being reduced, high temperature (550°C) PE nitride becomes more attractive due to its lower H content and enhanced etch selectivity to oxide. Its application can be found as the etch stop layer

in self-aligned contact and borderless contact etch processes, as well as sidewall spacers, with minimal metal contamination.

13.4.1.3 Silicon Oxynitride

One of the interesting features of PECVD films is the possibility of changing film composition continuously from oxide to nitride by varying gas flow. With the addition of N_2O into the SiH_4/NH_3 mixture in a PECVD reactor, the properties of the deposited film changes continuously from that of nitride to that of oxide. The resulting films are called silicon oxynitride (SiO_xN_y); these films have improved stability, reduced H impurity and better cracking resistance with low film stress. They are typically used for intermetal level dielectric and final passivation layers. Two-layer films of PECVD silicon oxynitride and silicon nitride have been used for planarizing multilevel interconnects by utilizing the differential etch rates of the two materials [57]. The enhanced interface properties between $Si–SiO_xN_y$ enabled direct film deposition as a gate dielectric [58] due to its impermeability to dopant diffusion and low electron trapping. As STI is becoming widely used as a replacement for LOCOS, silicon oxynitride also finds an application as the liner for trench isolation, with an improved process window, as well as being an effective O_2 diffusion barrier with resistance to hot phosphoric and HF acids [59].

13.4.1.3.1 Dielectric Anti-Reflection Coatings

Recently, a new application for silicon oxynitride is emerging as a DARC for sub-0.1 μm poly-Si gates as well as Al patterning using conventional DUV lithography. The SiO_xN_y refractive indices, both the real (n) and imaginary (k) parts, can be tuned over a wide range by varying its composition.

Dielectric anti-reflective coating is a dielectric material that has begun to replace spin-on organic arc layers for use in photolithography steps. Anti-reflective coatings are designed to limit reflections from a substrate during the photolithography steps which would otherwise interfere with the patterning process. The process of patterning photoresist on wafers is required in order to set up a mask for the etch or implant steps.

The process for patterning a wafer requires:

- Depositing the DARC layer onto the surface to be patterned;
- Spinning the photoresist layer onto the wafer;
- Exposing the photoresist to DUV light through a mask;
- Developing away the exposed areas of photoresist.

It is during the exposure step performed using a DUV stepper that reflections have to be controlled by the DARC layer on the surface of the wafer. Dielectric anti-reflective coating layers become more critical as the wavelength of light is reduced to 248 nm and eventually 193 nm as reflections become worse. It is also apparent that the dielectric arc layers have become more useful at these wavelengths since they have excellent CD control capability and can easily control reflections from any surface.

The DARC layer is deposited in the same way as is silicon nitride or oxide but the chemistry is such that the material becomes a combination of SiO_2 and SiN known as silicon oxynitride (SiON).

$$SiH_4(g) + N_2O(g) + He(g) \rightarrow SiON(s)$$

The Helium is used as a dilutent and improves overall uniformity and stability of the film.

13.4.1.3.1.1 Reflectivity Control

With DARC the mechanism for reflectivity control is primarily phase shift cancellation. This means that the reflected light is a combination of direct reflections and reflections that have gone through a half wavelength phase shift, resulting in destructive interference to the combined reflected light which then cancels out (Figure 13.16).

Spin on ARCs rely on absorption to control reflections, which requires a much thicker layer. They are also planar in deposition and result in varying thickness over topography, which can result in CD variations over steps. Figure 13.17 shows SEMs of advanced <0.25 μm structures patterned with DARC.

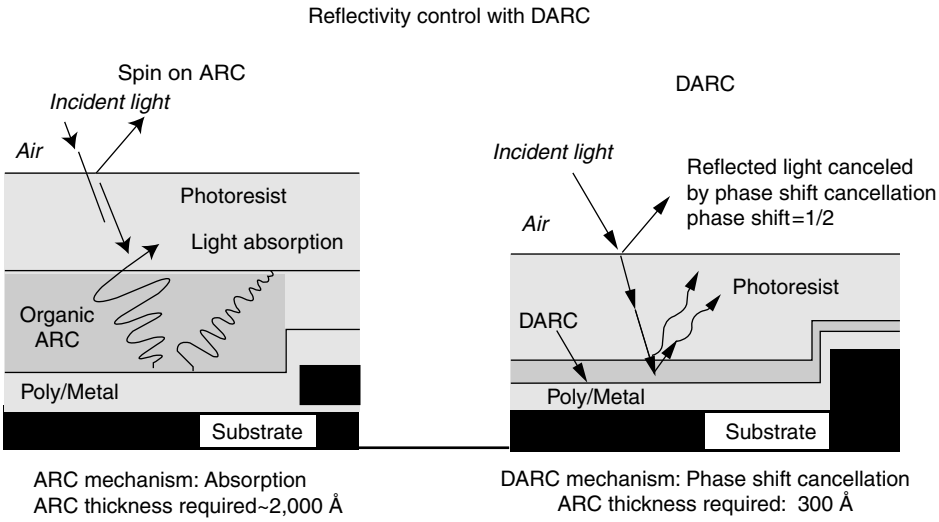


FIGURE 13.16 Reflectivity control: dielectric anti-reflective coating (DARC) vs. spin-on.

One of the measures customers use for ARCs in general is the amount of CD Swing measured as a percentage change as photoresist thickness is changed (Figure 13.18). Good CD Swing control is considered $<3\%$.

13.4.1.3.1.2 Integration of DARC

The main differences in integrating a DARC layer are that organic ARCs are always removed after use in photoresist strippers, while DARC material is normally left in or removed during etch or CMP steps.

13.4.1.3.1.3 DUV Photoresist Footing

With single-layer ARCs, the SiON material can react with the photoresist forming, which is known as a photoresist foot (Figure 13.19). Photoresist footing causes CD control issues. With DARC, footing can be

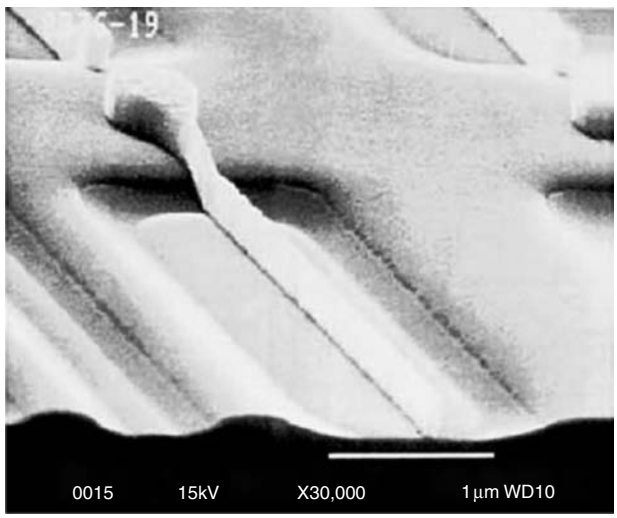


FIGURE 13.17 SEMS of advanced <0.25 μm structures patterned with DARC.

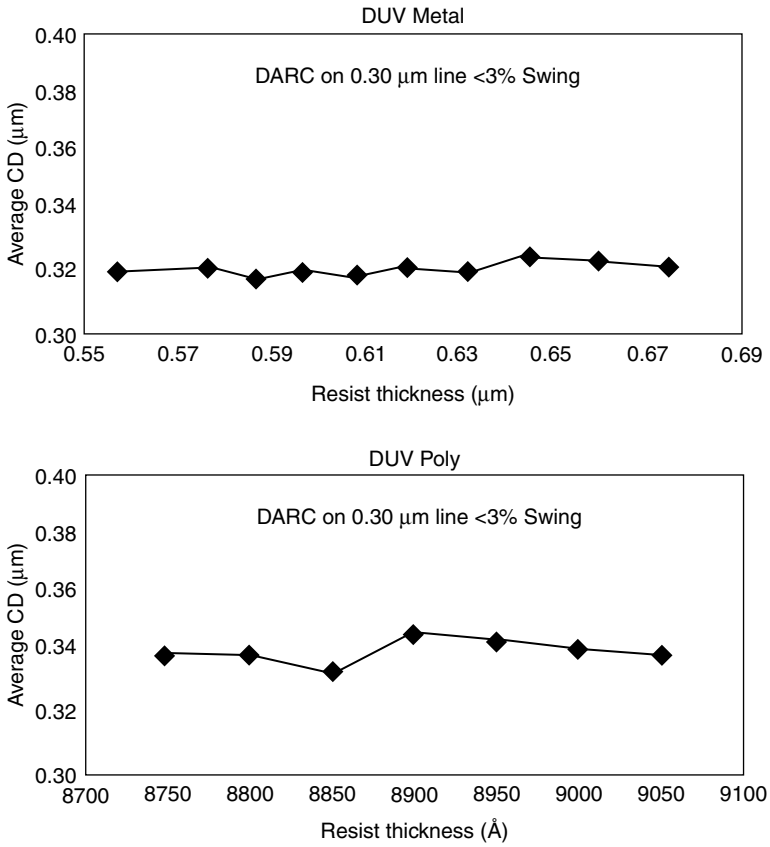


FIGURE 13.18 Critical dimension swing data for DARC over polysilicon and aluminum.

avoided by depositing an oxide cap layer that is inert and will not react with the photoresist layer. The oxide layer has no impact on the reflectance control.

13.4.1.4 Low *k* Dielectrics

The demand for low dielectric constant (low *k*) films in multi-level IMD applications is fast emerging and is expected to grow rapidly. As semiconductor device geometries shrink to 0.25 μm and below, the parasitic capacitance in the IMD between metal lines becomes more and more important in terms of RC time delay in device switching. In addition, the inter-metal capacitance needs to be reduced to minimized coupling between increasingly closely spaced metal lines, which leads to crosstalk, and to minimize power dissipation in dense device structures.

For the lowest possible capacitance between metal lines, vacuum is the naturally best medium. However, vacuum cannot provide mechanical protection and support to the device interconnect metallization. SiO_2 has been the primary material used for IMD electrical insulation in multi-level interconnects. The choice of SiO_2 was based on its good dielectric and mechanical strength, as well as, the ease of processing. However, SiO_2, with a dielectric constant ranging from 3.9 to 4.5 depending on formation methods, is not believed to be applicable in devices with geometries below about 0.18 μm because of its capacitance limitation. Various new materials—Si-based, C-based or a combination of both—have been mentioned as viable low *k* dielectrics of the future because of their dielectric constants from 3.7 to below 2.0 [60–64].

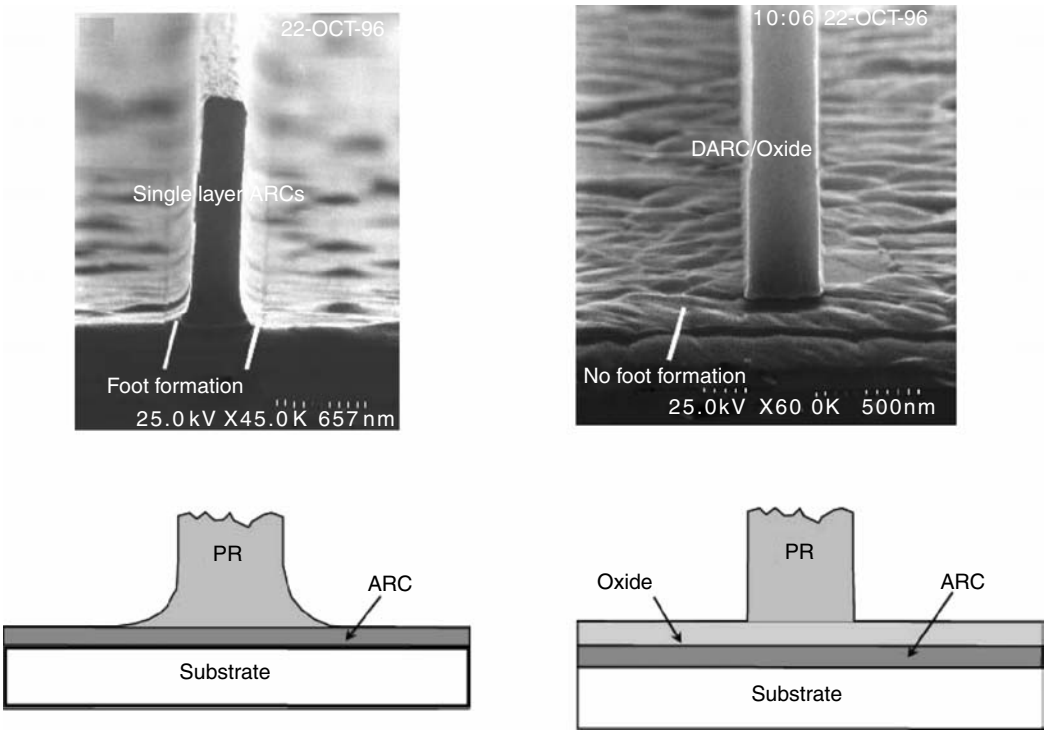


FIGURE 13.19 Photoresist footing mechanism.

13.4.1.4.1 CVD vs. Spin Coating

The two main deposition techniques for current low *k* materials are CVD and spin coating. Low *k* dielectric development started with chemical suppliers who engineered new chemicals with the desirable low dielectric constant. These chemicals were mostly "spin on dielectric" (SOD) where the dielectrics were dissolved in solvents and spun on the wafers, followed by baking and furnace curing for solvent removal and film densification. These chemicals are proprietary and expensive.

Besides its generally lower material cost and the advantages of using (typically) non-proprietary precursors, CVD techniques also offer considerable process flexibility. Bulk film and interface film properties can be easily altered in CVD processes by the adjustment of process gas flow ratio or other process parameters, while SOD properties can be changed only by modifying the precursor chemistry.

There are a variety of low *k* materials, CVD or spin on, currently available in the market that are Si and/or C-based, with very different characteristics.

From Table 13.6, C-based materials generally have lower *k* values, but the choice of low *k* materials also depends on other material characteristics which can actually favor Si-based materials that have higher dielectric constants. Generally, the properties of and process integration issues related to Si- and C-based low *k* materials are very different. Table 13.6 summarizes their typical differences in terms of materials properties and integration issues. Si-based materials usually have higher thermal stability and hardness compared to C-based materials, but Si-based materials tend to be more prone to moisture absorption.

With regard to integration issues, the Si-based materials are much more compatible to existing device manufacturing process flows. Adhesion to silicon, silicon oxides and nitride, as well as Al and Cu metallization, obviously is better in Si-based materials. Also, the familiar F-based etching chemistry can easily etch Si-based materials; whereas, O_2-based etching chemistry for C-based materials does not have

TABLE 13.6 Si-Based vs. C-Based Materials

	Si Based	C Based
Examples	Black Diamond, FSG, HSQ, Xerogel	PAE, BCB, SiLK
Material Properties		
Thermal stability	>425°C	350°C–450°C
Hardness	Good	10×less
Moisture absorption	Sensitive to moisture	Typically hydrophobic
Integration		
Adhesion	Good	Good, W/adhesion layer
Chemical–mechanical polishing compatibility	Dependent on deposition method	No direct polish
Etch compatibility	C_xF_y	O_2

as good a selectivity over organic photoresist. Probably the most important of all the possible integration issues is the CMP compatibility. Again, because of their greater hardness, the Si-based materials are more compatible with CMP processing, including direct polishing of the low *k* materials or supporting Cu polishing in the Damascene scheme.

13.4.1.4.2 Silicon-Based Low k Films

13.4.1.4.2.1 Fluorine-Doped Silicate Glass

One of the most straightforward ways to reduce the dielectric constant of IMD films is to dope SiO_2 with F in conventional CVD processes and equipment. Fluorine is one of the most electronegative elements. Therefore, F in the silicate network would tie up electron density around itself, making the overall film less polarizable, and hence reducing the dielectric constant. Figure 13.20 shows the molecular building blocks of Fluorine-doped silicate glass (FSG). Fluorine-doped silicate glass can be deposited with either SiH_4 or TEOS chemistry in conventional parallel plate PECVD reactors, or it can be deposited in HDP CVD reactors. The HDP approach has the advantage of giving significantly better gap-filling because of the simultaneous deposition and sputtering mechanism [65,66]. The fluorine source can be CF_4, C_2F_6, NF_3, SiF_4 or other organo-silicate sources [67,68]. Usually, SiF_4 or organo-silicate F sources that have Si–F bonds already present are preferred. This is because the incorporation of F into the silicate glass network using these precursors does not require breaking the strong Si–F bond, whose bond strength is roughly 135 kcal/mol, while with CF_4, C_2F_6, or NF_3 for example as precursors, F incorporation is through $F^{\cdot}$ and/or $CF^{\cdot}$ radical reactions which may result in free or loosely bonded F in the film.[9]

There is a fundamental limit to the amount of F incorporation in SiO_2 before the film becomes unstable. Theoretically, a large amount of F can be added to reduce the dielectric constant of FSG to 3.0 or even slightly below. However, the species SiF_2O is known to be volatile, meaning that FSG films will be

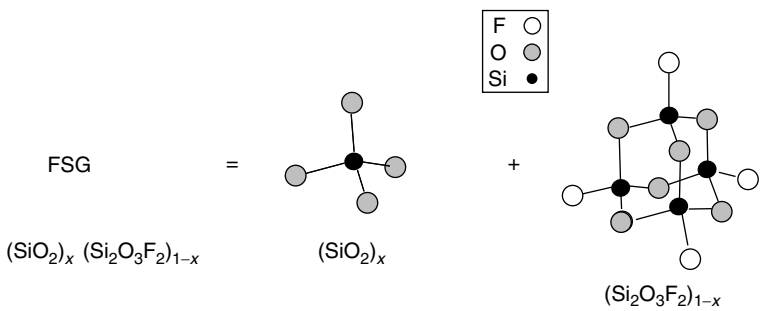


FIGURE 13.20 Molecular structure of fluorine-doped silicate glass materials.

unstable with F outgassing at high F concentration. Stable FSG films have been reported with dielectric constants down to about 3.5–3.7. Further reduction below this range will have trade-offs with reduced thermal stability.

13.4.1.4.2.2 Silsesquioxane

Another group of low *k* materials that has been used in manufacturing is Silsesquioxane. These materials were first manufactured for spin-on applications but can be deposited by CVD. It is basically a silicate material that contains one terminating group per Si. This terminating group can be H or other organic groups such as methyl or phenyl. The resulting silicate network will be distorted and forms 3D "caged" or "ladder" networks, as shown in Figure 13.21 with hydrogen silsesquioxane (HSQ) as an example that creates extra porosity for low dielectric constant. The effective *k* of HSQ has been reported between 2.9 and 3.9 [69,70]. If an organic group is used, the C content can further reduce the dielectric constant. Methyl silsesquioxane (MSQ) has been reported to have an atomic C content of ~25% and an effective *k* down to 2.5 [71,72].

13.4.1.4.2.3 Flow-Fill CVD Film

The motivation of developing a CVD alternative to spin-on silsesquioxane is primarily chemical cost. Spin-on precursors are proprietary and expensive, while the CVD processes use readily available and relatively inexpensive gas sources such as silane, methyl silane or phenyl silane and hydrogen peroxide. One possible CVD approach to depositing silsesquioxane materials is the flow-fill technology which involves the reaction of organo-silane precursors with H_2O_2 to form a silanol gel on a wafer usually cooled to around 0°C, and then followed by curing in a separate chamber or furnace to drive off the hydroxyl and leave behind a porous structure to achieve low dielectric constant [73]. Figure 13.22 shows the process flow of this CVD approach.

The key to obtaining the lowest dielectric constant for this kind of CVD low *k* material is to carefully optimize the curing process to completely drive off the hydroxyl while minimizing film shrinkage. The dielectric constant achieved by CVD is similar to that by spin coating. However, the CVD materials have lower C content, 8–10 at.% of C, and lower film density (higher porosity) compared to spin-on MSQ. Lower C content by itself will result in higher *k* but the lower film density will bring the *k* back to the same level with MSQ. This lower C content of the CVD materials is a definite advantage for process integration because it brings the materials closer to SiO_2.

Since the deposited silanol gel behaves like a viscous liquid, the gap-filling and degree of planarization over topography are also similar to the spin-on materials. Figure 13.23 shows the liquid-like bottom-up gap-fill of the CVD silsesquioxane materials using a sub-0.2 μm gap, and the planarization capability over a wide field.

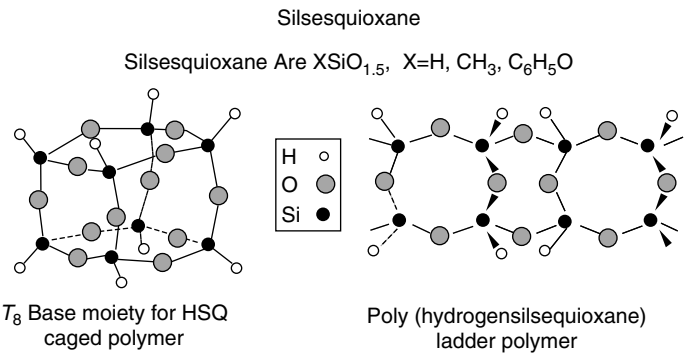


FIGURE 13.21 Molecular structure of silsesquioxane materials.

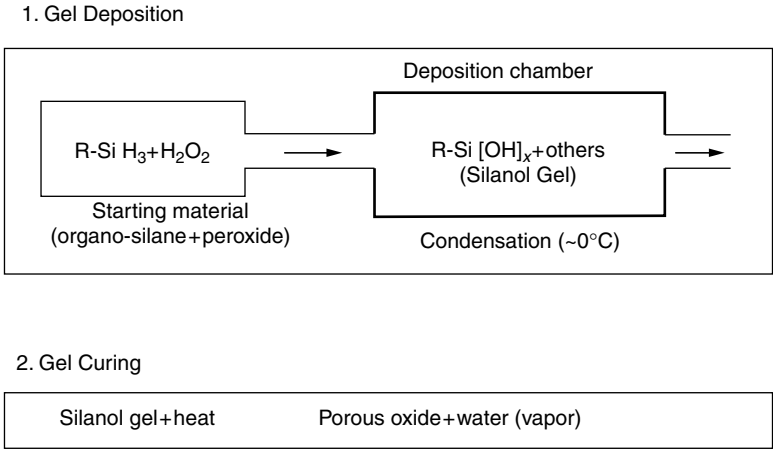


FIGURE 13.22 CVD approach to deposit silsesquioxane materials.

13.4.1.4.2.4 Black Diamond CVD Film

Another way to deposit SiO_2-like low-density material involves using a near room temperature PECVD technique with the same organo-silane precursors, but gaseous oxidizers like O_2 or N_2O instead of H_2O_2 [74,75]. This kind of deposition, called "black diamond" by one manufacturer, operates in a process regime that maximizes the sticking coefficient of the gas phase reactants. As a result, the as-deposited film already has a silsesquioxane type of porous structure. Since the deposition is at near room temperature, curing is still necessary to outgas the trapped reaction byproduct from the bulk film.

Compared with the organo-silane and H_2O_2-deposited CVD low k films, these films have material composition similar to SiO_2 but otherwise differ significantly in gap-filling ability, process simplicity and ability to extend to lower dielectric constant. Combining porosity and low C concentration of 8–10 at.% gives a dielectric constant of about 2.7. However, the high sticking coefficients of the gaseous reactants result in poor step coverage and gap-fill capability. Therefore, this film can be used as an IMD only in a damascene scheme.

Such processes use commercially available non-proprietary gaseous precursors and do not require heated plumbing and special isolated storage and delivery systems for the intrinsically unstable H_2O_2. A conventional single-wafer parallel plate-type chamber can be used; these chambers have been the

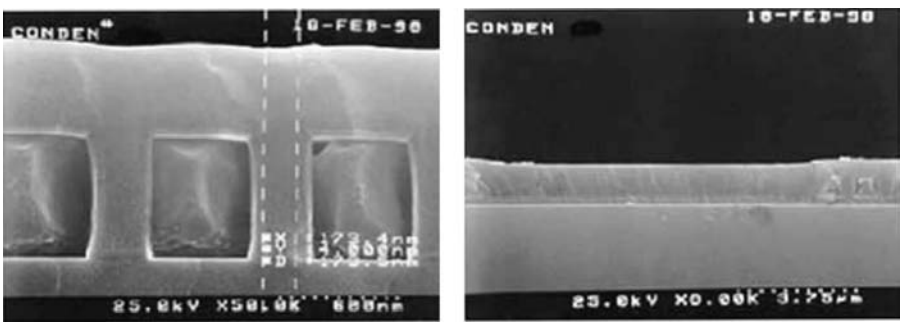


FIGURE 13.23 Liquid like gap-filling and planarization capability of CVD silsesquioxane materials.

mainstream hardware for IMDS used in high-volume manufacturing for more than a decade. This provides both process and hardware simplicity which will directly translate into low cost of ownership.

The ability to extend the dielectric constant of a type of low *k* material for more than one generation is a key requirement. Any new material for IMD needs to be evaluated vigorously not only for the intrinsic materials characteristics, but also for its compatibility to all the downstream processing during integration. For example, some soft organic spin-on materials show good static adhesion properties to metallization but after undergoing typical abrasion of CMP processing will start to delaminate. If a low *k* material is used in only one generation, these time-consuming and costly evaluations will be needed for consecutive generations of technology. Due to its ability to independently use process parameters to adjust the sticking coefficients of the gaseous reactants, and hence the degree of porosity in the film, without changing the materials composition, the technology described in this section can lower the dielectric constant of the film to close to 2.0 with increasing porosity while maintaining the SiO_2-like material characteristics.

13.4.1.4.2.5 Carbon-Based Low k Films

Carbon-based low *k* materials that do not contain any Si can be easily deposited by CVD techniques. These materials include PECVD or HDP amorphous carbon (a-C) and amorphous fluorocarbon (a-FC) [76,77], and also diamond-like carbon (DLC) and fluorocarbon (FDLC) [79,80]. All of these films use CH_4 as the main C source; in the case of a fluorinated version, CF_4, C_2F_6, C_3F_8 etc., are used for the F source. The film growth mechanism is similar to PECVD SiO_2. Again, the HDP approach gives significantly better gap-filling capability. The a-C and DLC films have dielectric constant ranging from 2.7 to close to 4.0 depending on molecular structure and H composition. Adding F to these materials can reduce the dielectric constant down to 2.5 or slightly below. Like SiO_2, the addition of F will likely degrade the thermal stability of the film.

13.4.1.4.2.6 The Parylene Family

One of the most promising types of carbon-based CVD low *k* materials is the parylene-based polymers. Parylene polymers are linear chains of phenyl groups isolated by ethyl groups at the para positions. Figure 13.24 shows the different types of parylene polymers.

The most efficient way to deposit these polymers is by the pyrolysis of the parylene dimer into the reactive di-radical/xylylene isomer and followed by condensation and polymerization at room temperature or lower [81,82]. Figure 13.25 shows the reaction pathway of this process. The phenyl

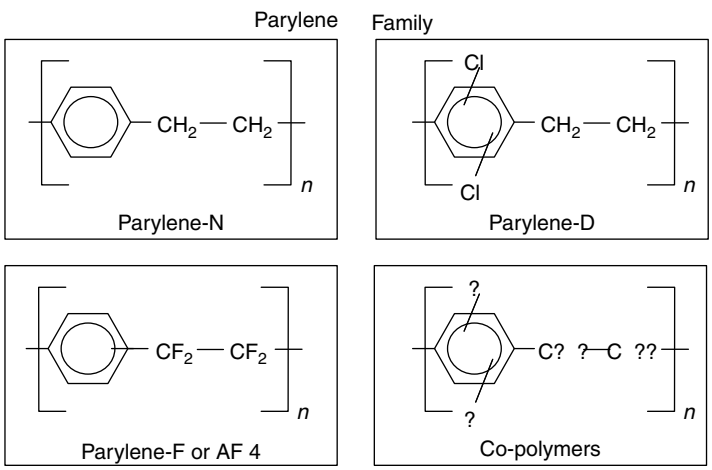


FIGURE 13.24 Different types of parylene polymers.

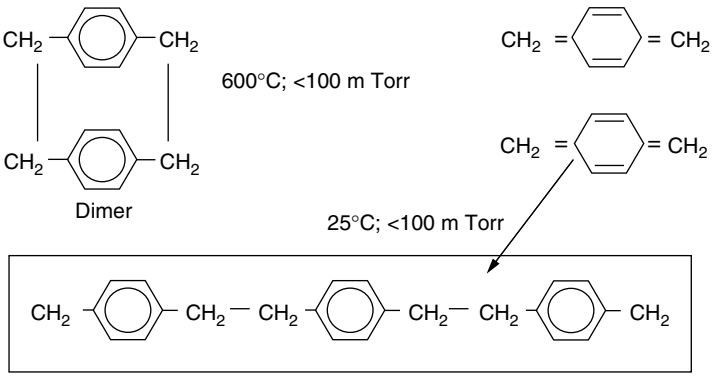


FIGURE 13.25 Reaction pathway of parylene polymerization.

group is known to have very evenly distributed electron density and is difficult to polarize. Hence, the dielectric constants of these polymers are low, at 2.7 or below. The parylene-N version has the highest dielectric constant at about 2.7. Doping parylene-N with other electronegative components such as F or Cl can reduce the dielectric constant to below 2.5. Parylene-F is known to have a dielectric constant between 2.2 and 2.3. However, F is always a problem with film stability and adhesion, and this may not be a viable option for IMD applications.

A different option to further reduce the dielectric constant as well as improve other film properties of parylene-N is through co-polymerization with a suitable co-monomer. The choice of the co-monomer determines the film properties of the co-polymers. The desirable co-monomers should be able to promote cross-linking to form a 3D polymer that can improve the thermal stability, as well as, incorporate more porosity to reduce the dielectric constant. In this regard, co-monomers that contain electron-donating molecular fragments such as the C=C bond can be suitable. In addition, the co-monomer may contain the silicate fragment which can increase the hardness and improve the adhesion of the co-polymer to other Si-based films or substrates.

The advantages of CVD co-polymerization are process flexibility and low cost. With a suitable co-monomer, one can easily adjust the amount of co-monomer in the process to tune the bulk film properties or selectively alter only the interface layer to promote adhesion to one particular deposition surface. As in most CVD processes, the precursors including parylene-N dimer and various co-monomers are readily available and inexpensive. An example of a CVD process flow for the parylene based co-polymerization can be found in Figure 13.26. This kind of co-polymer has already demonstrated dielectric constant values of 2.0. With better choice of co-monomers, the dielectric constant will likely be reduced to below 2.0 with this technique.

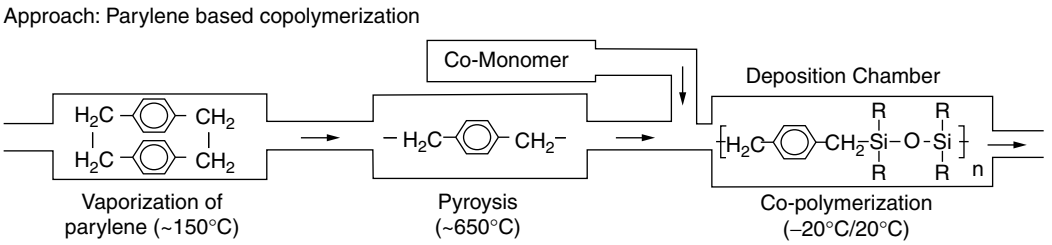


FIGURE 13.26 CVD process flow for parylene based co-polymerization.

13.4.1.4.2.7 Directions of Low k Technology

Even though low dielectric constant is the primary goal for low *k* materials development, the ease of process integration, and not how low the *k*-value is, ultimately determines the final choice of a low *k* dielectric used in device fabrication. Logic device manufacturers are the main drivers for low RC requirements. Sub 0.18 μm logic devices require Cu metallization, not only for lower wiring resistance but also yield improvement and cost saving resulting from the reduction of metal levels. This means that low *k* materials will be primarily integrated with Cu in the damascene process scheme.

Typical damascene process steps involve trench and via etching of the dielectric as well as Cu and barrier fill and CMP directly on the dielectric. Etching processes will have to be developed for each type of low *k* materials. The dielectric will also have to have good adhesion to Cu and barrier metals, and be mechanically strong enough to withstand the CMP processes. With all these considerations, low cost Si based materials which are reasonably close to SiO_2 in their film properties will likely be the choice for low *k* dielectric going from *k* of 4.0 down to 2.5. These materials, such as FSG and black diamond films, will substantially reduce process development time and resources for etch and CMP processes. In addition, their SiO_2-like properties will provide the necessary adhesion and mechanical strength for the Cu damascene process scheme.

Going down to the technology node where *k* of 2.5 or below is required, there appear to be two possible migration paths. The first one is to continue with Si-based materials and introduce additional porosity into the film to reduce k. The extendibility of black diamond-type films to a *k* value close to 2.0 makes it an attractive candidate. The advantage of this approach is the similar process integration schemes that can be extended from previous generations. Possible disadvantages include lower mechanical strength and moisture absorption due to the porosity.

The second path is to switch to carbon-based organic materials which generally have lower *k* than Si-based materials. A number of process integration issues with these materials will have to be solved. Development in materials with *k* below 2.5 is still preliminary at this point. It appears likely that the industry will stay with Si-based low *k* materials as long as possible until the point when the *k*-value cannot meet the RC requirements. This transition point may never happen if Si-based materials can prove extendibility to $k < 2.0$ or the integration issues of C-based materials cannot be resolved in cost effective ways.

13.4.1.5 High *k* Dielectric

13.4.1.5.1 Ta_2O_5 for DRAM Applications

The density of the metal-oxide semiconductor (transistor) DRAM capacitor chip is continuously increasing. At the same time the capacitance per unit cell is also increasing. The dielectric of choice for the present generation of DRAM devices is the composite of Si_3N_4 and SiO_2 (NO or ONO) films. The dielectric thickness for the "NO" dielectric has already reached the limit of tunneling current, hence for higher density DRAMs a higher *k* dielectric material will have to be used. Ta_2O_5 is a CVD high *k* dielectric that can be integrated into the current device flow of "NO" dielectric with minimal impact on the device integration flow [83,84].

The capacitance specification for the DRAM application is quoted in terms of equivalent thickness of SiO_2. The current specification for 0.18 μm device geometry is T_{ox} of 30 Å and $J_c < 1 \times 10^{-8}$ Å/cm^2 at V_c of 1.25 V. For high density DRAM applications, the cell structures are 3D crown or cylindrical structures with feature sizes of < 0.2 μm and A/R $> 6{:}1$. The surface area of these cells is enhanced by having HSG or RSP along the walls of the cell. Thus, the dielectric film must have a very good step coverage to meet the leakage current and effective oxide thickness specification.

13.4.1.5.2 Ta Precursors

The Ta_2O_5 films are deposited using a thermal CVD process with metal-organic precursors. Two commonly found precursors of Ta, tantalum ethoxide (TAETO) and tantalum tetraethoxy dimethylaminoethoxide (TATDMAE), have a vapor pressure several orders of magnitude less than other metal-organic liquids used commonly in the semiconductor industry, such as TEOS (Figure 13.27).

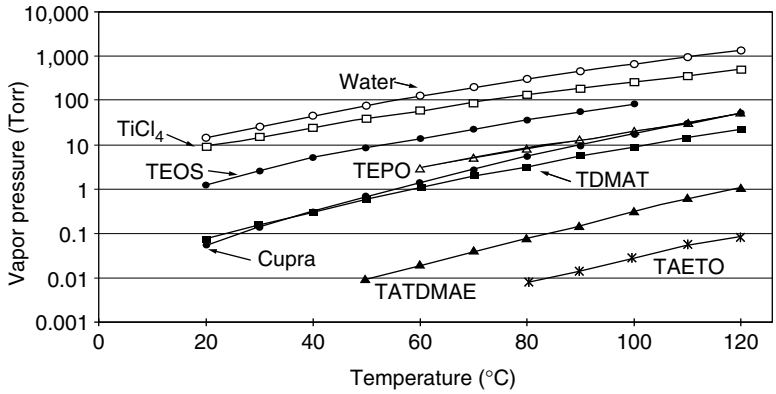


FIGURE 13.27 Vapor pressure as a function of temperature for several liquids commonly used in the semiconductor industry. The vapor pressure of TAETO and TATDMAE are the lowest of the chemicals listed.

The freezing point of pure TATDMAE is 9°C whereas that of pure TAETO is 18°C. Thus, the challenging task of any Ta_2O_5 deposition chamber is delivering a fully vaporized precursor to the wafer surface without any condensation or decomposition anywhere in the liquid delivery system. It is impractical to use pure TAETO since in order to prevent condensation of the liquid, the ampule and the entire length of the line from the ampule cabinet to the chamber has to be heated to high temperatures. To lower the freezing point of TAETO, it is usually mixed with ethanol (EtOH). Tantalum ethoxide with 2.5% EtOH has a freezing point of <10°C and poses fewer challenges than pure TAETO.

13.4.1.5.3 Deposition Process

The deposition rate is sensitive to temperature and pressure as shown in the Arhenius plots of Figure 13.28. The temperature range of 400°C–485°C represents a reaction rate limited regime with the activation energy of the deposition process being 1.3 eV. The deposition rate doubles every 25°C. Above 485°C the process transitions into a mass flow limited regime.

For good step coverage the deposition process has to be carried out in the reaction rate limited region. The amount of Ta flow to be used is determined by the Ta flow vs. deposition rate curve. This curve shows a knee with the deposition rate remaining constant after a threshold Ta flow. This regime is called

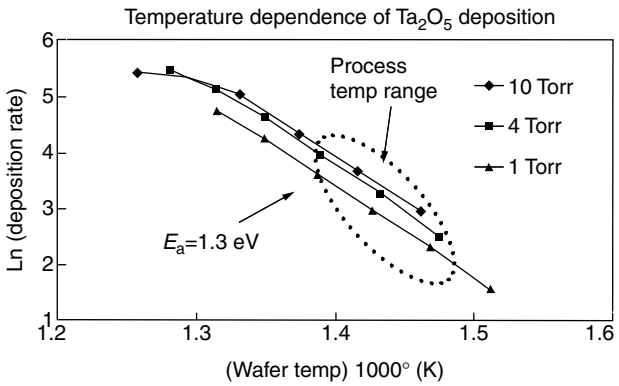


FIGURE 13.28 Temperature and pressure dependence of Ta_2O_5 deposition rate. The reaction rate limited regime is between 400 and 485°C.

the kinetically saturated regime. Liquid flow in the kinetically saturated regime has to be used to get good step coverage for device geometry <0.18 μm. For TAETO and TATDMAE this flow has been determined to be 35 mg/min.

Typical deposition process parameters are:

Wafer temp	450°C
Chamber pressure	4 Torr
O_2+N_2 process	500 + 1500 sccm
Precursor flow	35 mg/min
N_2 carry	600 sccm
Deposition rate	30 Å/min (for TAETO 2.5% EtOH) or 40 Å/min (for TATDMAE)

13.4.1.5.4 Step Coverage

Figure 13.29 shows step coverage results for this process with the TATDMAE precursor. The cross-section TEM for a 70 Å film on a 15:1 A/R trench structure with opening size of 0.3 μm using a 450°C, 4 Torr deposition process resulted in a sidewall and bottom coverage of $>90\%$ (Figure 13.29a). The cross-section TEM of the rough surface polysilicon substrate where 100 Å of Ta_2O_5 was deposited shows good conformality at the base of the grains as well as between the grains (Figure 13.29b).

13.4.1.5.5 Integration of Ta_2O_5 into Stack Capacitor

The TaO MOS stack capacitor has several interfaces (Figure 13.30). As the dielectric thickness decreases, the interfaces become important and the interface capacitance can lower the total capacitance of the stack.

As-deposited TaO film is oxygen deficient and resistive in nature. Oxygen annealing of this film is essential for it to act as an effective dielectric material (Figure 13.31).

13.4.1.5.6 Annealing of Ta_2O_5

The oxidation anneal is required to improve leakage characteristics of the stack capacitor. However, the TaO anneal has to be optimized to prevent Si diffusion and oxidation at the SiN/TaO interface. High-temperature annealing results in Si diffusion and the aggressive oxidizing environment leads to the formation of a SiO_2 layer between SiN and TaO (Figure 13.32).

The result of formation of this SiO_2 (low k) dielectric material at the interface is improved leakage current but a significant reduction in capacitance. Figure 13.33 shows the J–V curve for the sample annealed at 800°C in an N_2O ambient. Leakage currents are $<1\times10^{-8}$ Å/cm$_2$ at ±1.5 V. However

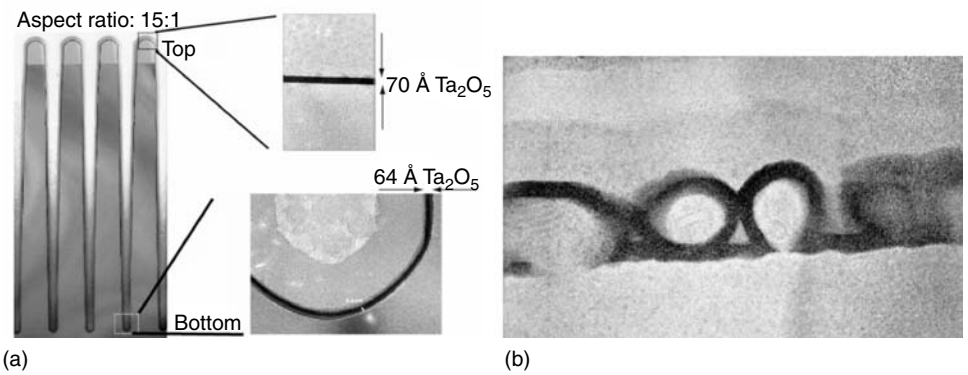


FIGURE 13.29 Step coverage of CVD Ta_2O_5 structures.

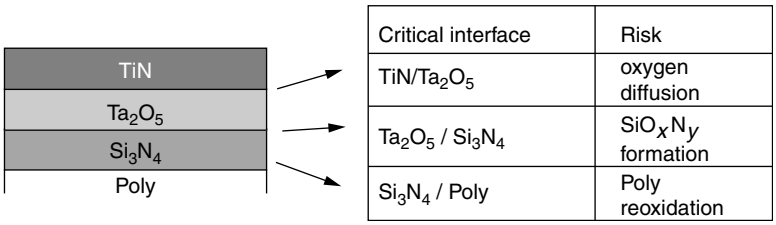


FIGURE 13.30 TaO metal-oxide semiconductor (transistor) stack capacitor showing different interfaces and possible problems posed at each interface.

the T_{ox} is 38 Å, which is significantly higher than the required specification of 30 Å for the typical 0.18 mm DRAM device geometry.

One of the reasons for the diffusion of Si through the SiN barrier layer is the quality of the SiN barrier layer. Figure 13.34 compares RT-SiN (Sample A) vs. CVD SiN (Sample B) for similar oxidation anneal of the TaO. While the capacitance of the stack capacitors in the two samples were similar, the leakage performance of the CVD SiN sample was better ($V_c \sim 1.2$ V) as compared to the RT-SiN ($V_c = 0.8$ V). This example illustrates that a good quality barrier layer can prevent the diffusion of Si.

Optimization of anneal is very critical to achieve a balance between lower leakage currents and high capacitance values for a Ta_2O_5 stack capacitor. Table 13.7 below gives the electrical performance of Ta_2O_5 MOS stack capacitor with various oxidation anneals. The bottom electrode was planar-doped polysilicon and the top electrode was PVD TiN. The critical voltage (V_c) is defined as the threshold voltage at which leakage current (J_c) is $< 1 \times 10^{-8}$ Å/cm^2.

13.4.1.5.7 Effect of Precursor on Electrical Properties

We have seen significant effects of the tantalum precursor on the electrical properties of MOS stack capacitors with tantalum as the high *k* dielectric material. The process integration flow to achieve similar electrical results with different precursors also must be optimized.

The vapor pressure of the TATDMAE precursor is an order of magnitude more than TAETO at any given temperature; hence from a hardware reliability standpoint it would be beneficial to use TATDMAE. However, Ta_2O_5 deposited using TATDMAE gives ~20% lower capacitance values than that deposited

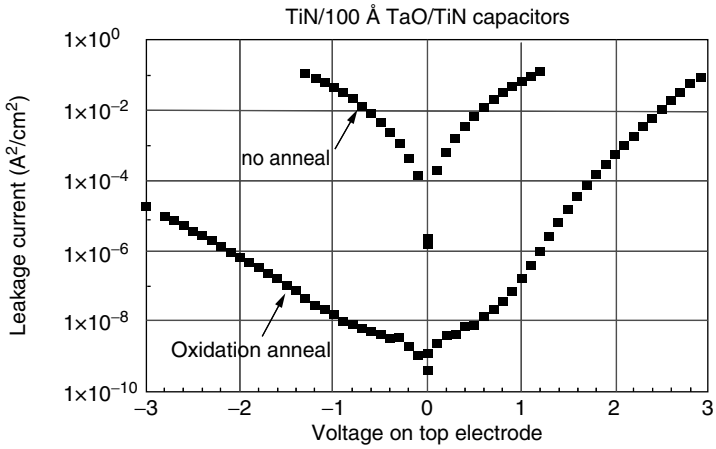


FIGURE 13.31 As-deposited TaO film.

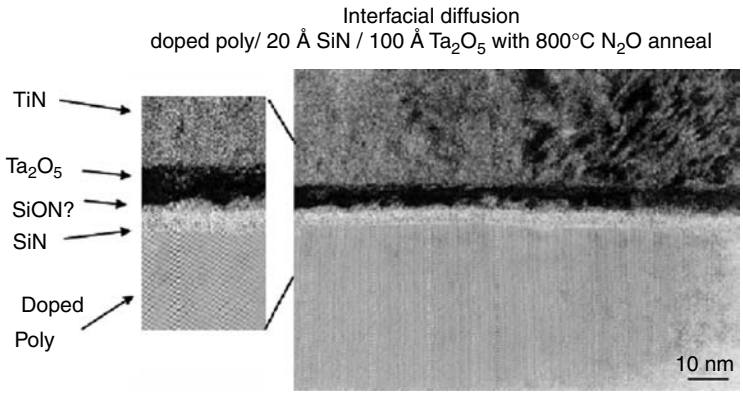


FIGURE 13.32 Cross-section TEM for an 800°C N_2O annealed TaO MOS stack capacitor showing interfacial diffusion resulting in the formation of SiO_2 at the SiN/TaO interface.

using TAETO for the exact same process integration flow (Figure 13.35). The leakage current appears similar for the two films; however, the effective oxide thickness for the film deposited using TATDMAE is ~7 Å greater than the one deposited using TAETO. Thus, the process where Ta_2O_5 is deposited using the TATDMAE precursor is not a "drop in" process; i.e., the users will have to modify their current process integration flow which they have optimized for TAETO in order to achieve similar electrical performance from the films deposited using TATDMAE.

Table 13.8 below compares the electrical performance for films deposited using the two different precursors and annealed using a low-temperature remote plasma anneal process. The annealing was performed at temperatures lower than the crystallization temperature of TaO.

Above data suggests that the capacitance of the films grown using TAETO are ~20% higher than those grown using TATDMAE.

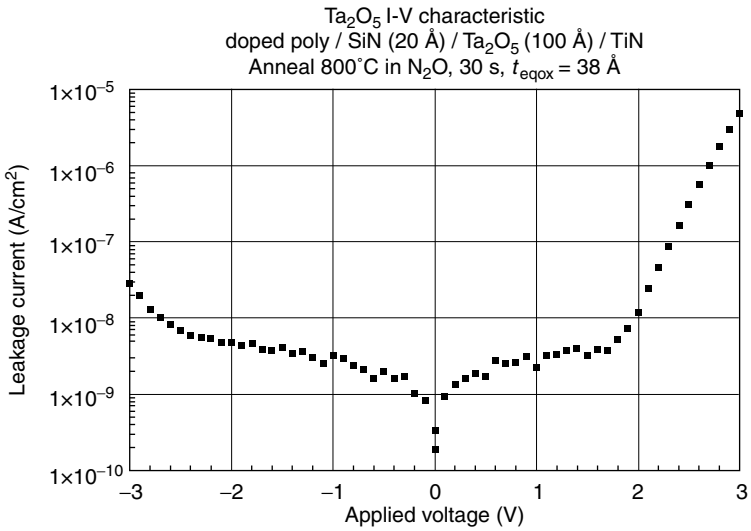


FIGURE 13.33 *J–V* characteristics of the Ta_2O_5 MOS stack capacitor after RT-N_2O anneal at 800°C.

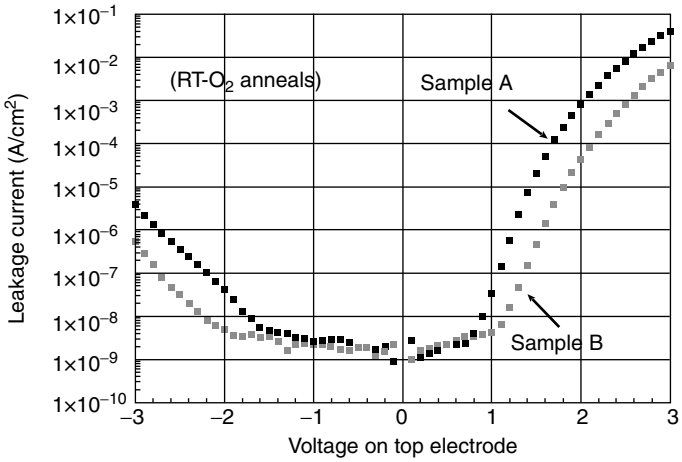


FIGURE 13.34 Sample A: RTN 20 Å/Ta_2O_5 100 Å + RT-O_2/TiN; Sample B: CVD SiN 20 Å/Ta_2O_5 100 Å + RT-O_2/TiN.

13.4.1.6 CVD of Barium Strontium Titanate

Barium strontium titanate (BST) is a leading candidate material for use as a capacitor dielectric in future generations of high-density DRAMs due to its high relative dielectric constant (high k) [85–87]. In bulk form, k can be as great as ~2000. Grown as a thin film (on the order of a few hundred Ångstroms), k is typically 200–400. Barium strontium titanate is a crystalline perovskite-structured oxide material with chemical formula $(Ba_xSr_{1-x})\ TiO_3$. Common compositions for DRAM use are $0.5 < x < 0.7$.

The specific dielectric properties requirements for BST in a DRAM cell are a function of the capacitor geometry. The minimum amount of charge that must be stored is 25 fF/cell (this limit is determined by the detection capability of the DRAM sense amplifier). Assuming a 3D storage node (post) that is $3f \times f \times h$, where f is the critical feature size and h is the storage node height, the minimum capacitance density (fF/mm^2) of the BST film needed to store 25 fF/cell can be calculated (Figure 13.36). For example, if BST is used in an $f = 0.15$ mm geometry with an aspect ratio of 2:1, the film must have a capacitance density of ~80 fF/mm^2. Additionally, film leakage and dielectric loss must be low $\sim 1 \times 10^{-8}$ Å/cm^2 and tan $d = 0.006$ [2], respectively. Since these material requirements will have to be met on 3D capacitor structures, good conformality (80%–100%) is needed as well.

13.4.1.6.1 Precursors

Most precursors for CVD BST are solids at room temperature and are dissolved in an organic solvent for delivery as a liquid to a vaporizer. The liquid is flash-vaporized (to minimize the time precursors spend at elevated temperatures and hence, minimize precursor decomposition) and transported in a gaseous state

TABLE 13.7 Electrical Performance of Ta_2O_5 MOS Stack Capacitors for Various Oxidation Anneals of the Dielectric Film

Anneal Condition	T_{eff}	V_c
475C RPO	34	2.0
800C O_2	30	1.25
750C N_2O	33	1.50
Crystallization + 475 RPO	25	1.15
Treatment of SiN/750 N_2O	30	1.25

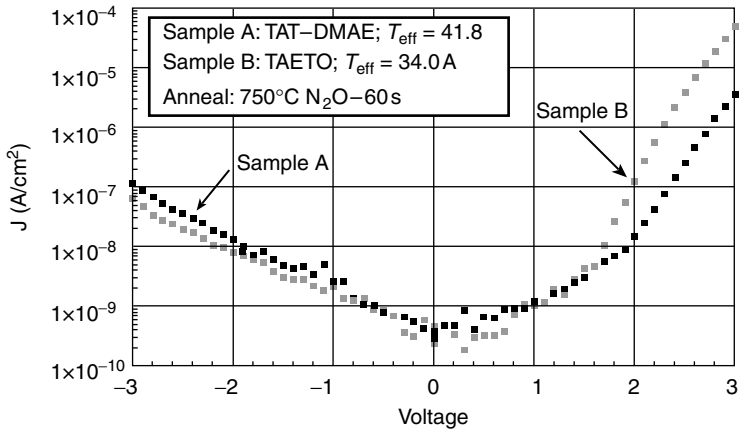


FIGURE 13.35 *J–V* and T_{ox} comparison between Ta_2O_5 films deposited using TAETO and TATDMAE.

to the CVD chamber. This liquid delivery enables the use of solid precursors; however, it chemically complicates the deposition process by introducing a solvent. Most of the published work on CVD BST uses precursors that are adducts of barium and strontium β-diketonates along with organometallic titanium compounds, primarily $Ba(THD)_2 \cdot$tetraglyme, $Ba(THD)_2 \cdot$PMDETA, $Sr(THD)_2 \cdot$tetraglyme, $Sr(THD)_2 \cdot$PMDETA, and $(THD)_2Ti(i\text{-Pr–O})_2$. These are typically dissolved in butyl acetate. Barium strontium titanate films of good quality (meeting or exceeding electrical requirements for DRAMs) have been produced using this precursor scheme at deposition wafer temperatures at 550°C. However, in this temperature range conformality is typically poor and may not be acceptable for 3D capacitor fabrication. Another set of precursors appears promising for lower temperature BST film growth. Non-adducted versions of the same barium and strontium β-diketonates, $(Ba(THD)_2)_4$ and $(Sr(THD)_2)_4$, along with $(THD)_2Ti(i\text{-Pr–O})_2$ are dissolved in tetrahydrofuran. Reports in the literature state that BST films with capacitance densities (planar capacitors) over 100 fF/mm^2 and leakage currents in the 10^{-8} A/cm^2 range can be grown at wafer temperatures of 480°C using these precursors [3]. Other work has demonstrated conformality of 100% at 480°C using the non-adducted precursors (Figure 13.37).

The reason for the difference in conformality between low and high temperatures is that different mechanisms are limiting the deposition process. Figure 13.38 shows an example of an Arrhenius plot of film deposition. Two distinct regimes are apparent with a crossover at ∼480°C. In the high temperature regime (where deposition rate is relatively constant with temperature), film growth is rate-limited by mass transport from the vapor phase to the film surface, whereas in the low temperature regime, film growth is limited by the chemical reaction rate on the surface. Films grown at high temperatures have

TABLE 13.8 Remote Plasma Oxidation of Amorphous TaO films Grown Using TAETO and TATDMAE

		TaO (Target 100 A)					
RPO		TAETO			TATDMAE		
Temperature/ Time	Crystalliztion	C_p (fF/μm^2)	J at 1.5 V	T_{eff}	C_p (fF/μm^2)	J at 1.5 V	T_{eff}
400°C; 2 min	No	9.58	4.10×10^{-9}	36	7.79	5.80×10^{-9}	44
500°C; 2 min	No	9.63	1.20×10^{-9}	36	7.85	1.40×10^{-9}	44
600°C; 2 min	No	9.58	6.00×10^{-9}	36	7.82	5.00×10^{-9}	44.2

Integration sequence: 90°C RT-NH_3 (20 A)/100 A TaO/RPO

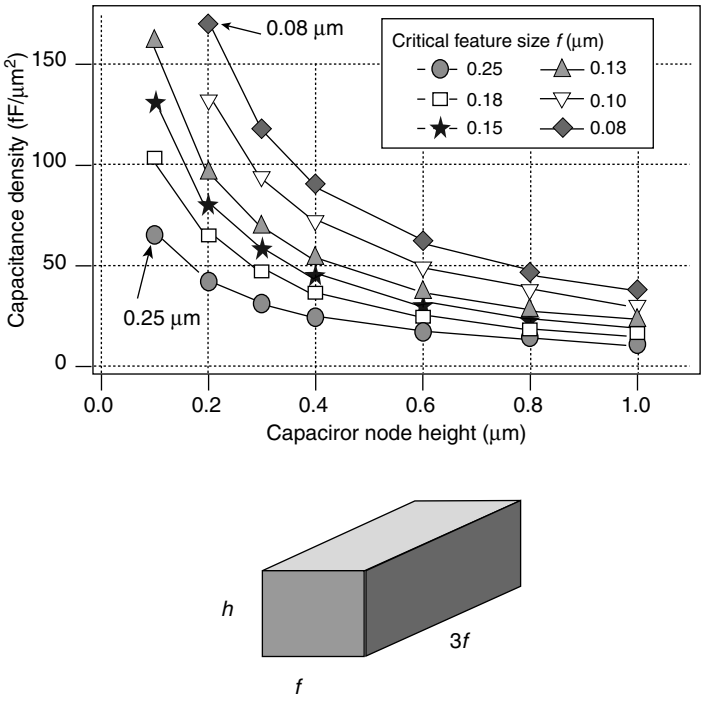


FIGURE 13.36 Minimum capacitance density requirements for BST films in a 25 fF/cell dynamic random access memory cell with post dimensions $3f \times f \times h$.

poorer conformality as precursor material reacts immediately upon arriving at the film surface and does not have time to diffuse into narrow patterned features before incorporating into the film, as it does in the surface reaction limited temperature regime. In BST deposition, a competition occurs between, requiring a high enough deposition temperature to ensure good crystallinity of the perovskite phase,

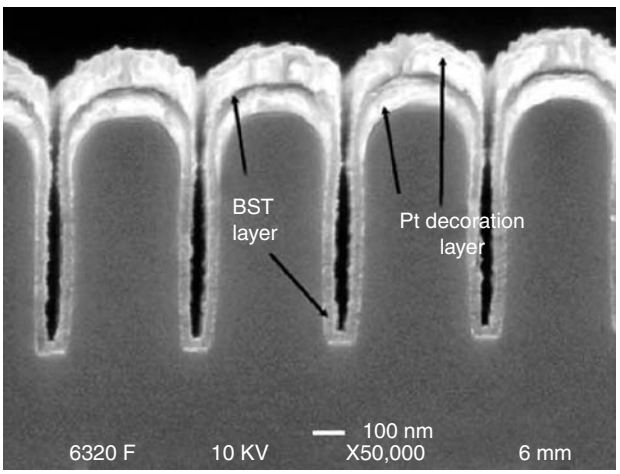


FIGURE 13.37 SEM micrograph of conformal BST deposition. Trench width is 0.13 μm; aspect ratio $\geq$6.5 to 1.

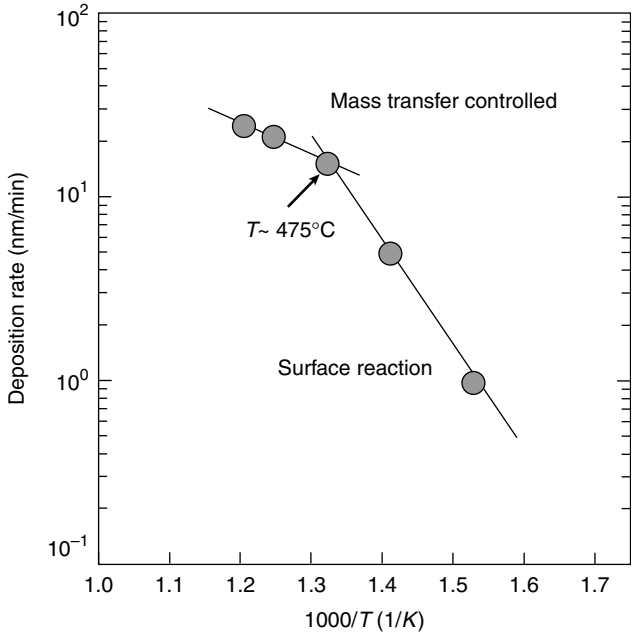


FIGURE 13.38 Arrhenius plot showing two regimes of rate limiting mechanisms for CVD BST deposition.

and requiring a low enough temperature to achieve adequate conformality. The adducted precursors form highly crystalline perovskite films at high temperatures, but these films deposited at low temperatures (surface reaction limited regime) are mostly amorphous. Non-adducted precursors, however, can produce perovskite BST films in the low temperature regime.

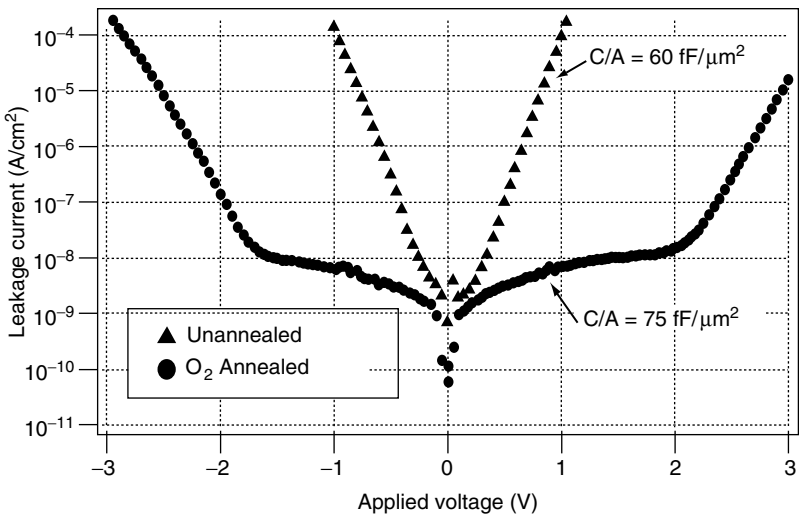


FIGURE 13.39 Leakage curves from unannealed and 550°C oxygen-annealed BST films deposited at 600°C. Labels on the leakage curves give the corresponding capacitance densities of the films.

After deposition, CVD BST films are typically annealed to improve the electrical properties. Annealing is usually done after top electrode deposition (particularly if the top electrode is deposited using a technique that bombards the BST surface with energetic species, such as PVD, the idea being to anneal out damage done to the BST by top electrode deposition), and reported anneal parameters range widely. Temperatures are reported from ~500 to 700°C, and both oxidizing (N_2O, O_2) and inert (N_2) ambient are reported to improve electrical properties. Typically, films grown at high temperatures (550°C) using the adducted precursors require lower annealing temperatures, around 550°C in oxygen is commonly used. Low temperature depositions require higher annealing temperatures, closer to 700°C.

The observed effect of annealing is mainly an improvement in the leakage characteristics of BST capacitors. Before annealing, films are leaky (Figure 13.39) although capacitance is relatively high. After annealing, leakage improves to an acceptable level, and capacitance increases by roughly 25%, although the improvement is not as dramatic as that seen in the leakage properties.

13.4.1.6.2 Microstructural Effects on Electrical Properties

One of the most important factors for controlling BST electrical properties is the amount of titanium in the film. Titanium content is referred to as a percentage of all cations in the film, i.e., Ti% = [Ti]/[(Ba) + (Sr) + (Ti)] × 100%. It has been shown that optimal properties (highest capacitance density) are obtained with slight excess of titanium, ~51% (compared to the stoichiometric perovskite composition of 50%). Above this amount, the dielectric constant decreases (Figure 13.40). Given the sensitivity of the electrical properties to the titanium content (in this example, dielectric constant drops ~40% from 51 to 53% titanium), control of titanium incorporation during deposition is critical. A promising result is the demonstration of a titanium composition self-matching phenomenon using the low temperature, non-adducted precursor process. Figure 13.41 shows that titanium incorporation is stable around 50% for titanium precursor flows ranging from 20 to 80 mg/min (factor of four variations in flow). This relative insensitivity of film titanium content to amount of available titanium precursor material could prove very useful in developing robust BST deposition processes for DRAM manufacturing.

Another important film characteristic that has great impact on the electrical properties is the surface roughness. Increased surface roughness correlates with lower dielectric constants and higher leakage currents. Roughness is often described by the term "haze," as rough films look hazy to the naked eye under bright light. Roughness can result from film deposition on a rough underlying layer, as well as

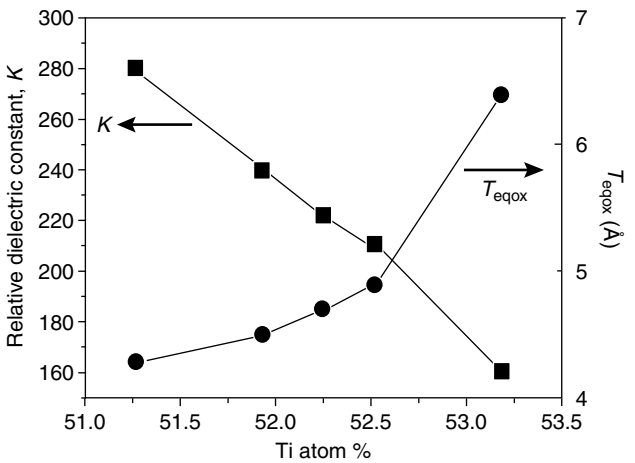


FIGURE 13.40 Dielectric properties variations (relative dielectric constant k and equivalent oxide thickness T_{eqox}) with the titanium percentage in ~280 Å thick BST films.

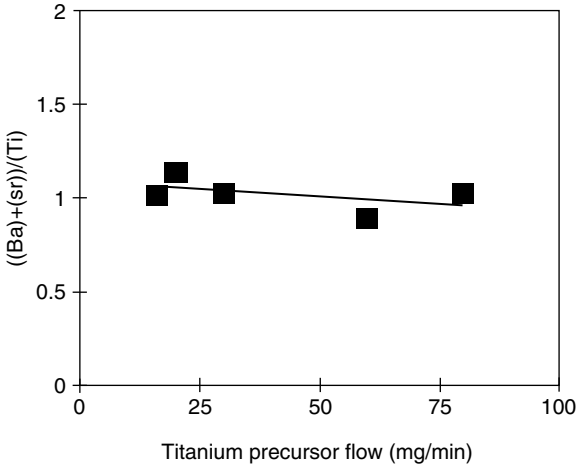


FIGURE 13.41 Cation ratios in BST films deposited using differing titanium precursor flow rates. Composition self-matching is seen at ~50% titanium.

from film growth occurring in an island mode. Roughness resulting from film growth mode is very sensitive to deposition process parameters, and parameters must be optimized to find a "haze-free" process window. Typically, less haze is seen at higher deposition temperatures, above 600°C.

13.4.1.6.3 BST Integration Issues

Integrating BST into silicon-based microelectronics brings with it a host of other issues. Barium strontium titanate is incompatible with conducting electrode materials commonly used in microelectronics today, materials such as titanium nitride, polysilicon, and aluminum. Platinum is the most commonly used electrode for BST in research and development as it is non-reactive with BST and forms an interface with BST that has favorable electrical properties. However, platinum integration presents some difficulties. Due to its chemically inert nature, it is difficult to dry-etch. It does form silicides and thus must be deposited on a diffusion barrier layer to prevent Si–Pt reactions. The barrier layer must also serve as an adhesion layer, as platinum has poor adhesion on most surfaces. This entire lower electrode structure must be oxidation resistant and thermally stable to withstand the high-temperature, oxidizing environments of BST deposition and annealing. Also, platinum catalyzes the decomposition of molecular hydrogen into atomic hydrogen. This is a concern in forming gas anneals during back-end processing, as BST electrical properties are degraded by exposure to hydrogen. Platinum electrodes can actually catalyze and enhance this degradation. Other electrode materials can be used as well, such as ruthenium metal, ruthenium oxide (RuO_2), iridium oxide (IrO_2), or conducting perovskite oxides such as $SrRuO_3$ or (La, Sr) CoO_3. However, none of these is standard in semiconductor processing, and all will require development for incorporation into DRAMs.

13.4.1.7 ALD High *k* Materials

13.4.1.7.1 Metal Oxide

SiO_2 has been used in Si-based microelectronic devices as gate dielectric material. As the device technology moves forward, the gate oxide thickness is scaled down aggressively to improve transistor performance. According to international technology roadmap for semiconductors, low power devices will require an equivalent oxide thickness (EOT) of $\leq$15 Å for 65 nm technology node, while high performance devices require an EOT of $\leq$10 Å. On the other hand, the tunneling current through gate dielectric increases exponentially by decreasing the gate oxide thickness. The tunneling current will

become detrimentally large by further reducing the thickness. Gate dielectric with higher *k* constant must be employed to increase gate dielectric thickness while maintaining the same gate capacitance.

The following summarizes the requirements for high *k* gate dielectric materials

- High relative permittivity (>10);
- Large conduction band offsets;
- High electric field strength;
- Low defect density (pin-hole free);
- Thermal stability (amorphous structure);
- No interaction with Si substrates;
- Good control of stoichiometry;
- Smooth surface morphology;
- Good thickness uniformity and thickness control.

High-*k* metal oxide films such as Al_2O_3, ZrO_2, HfO_2 and Y_2O_3 have been suggested to replace silicon oxide as gate dielectric materials. Ta_2O_5 and TiO_2 are not suitable for gate oxide application due to the instability in contact with Si substrate. Figure 13.42 summarizes the *k*-value and band gap offsets of high-*k* gate dielectric candidates.

Among these materials, ZrO_2 and HfO_2 are of the most interesting materials because both hafnium and zirconium oxides have relatively high dielectric constants and band gaps compared to the rest of materials. Thermodynamic considerations indicate that both zirconium and hafnium oxides should be stable in contact with a silicon surface. In the use of hafnium and zirconium oxides for gate dielectric applications, however, still many technology challenges exist. One of these is that interfacial layers are formed in contact with Si substrates, which significantly increase EOT. In addition, both ZrO_2 and HfO_2 show strong crystallization tendencies, which lead to rough interface morphology and potential of current leakage. Hafnium and zirconium oxides have been deposited at less than 200°C to suppress film crystallization. In this experiment, HfO_2 and ZrO_2 are deposited using tetrakis(dimethylamido)hafnium and tetrakis(dimethylamido)zirconium where H_2O is used as the oxidant. The results, however, revealed that crystallites are observed at deposition temperatures as low as 200°C for HfO_2 and 100°C for ZrO_2, respectively. Periodic introduction of a second oxide with amorphous structure such as aluminum oxide

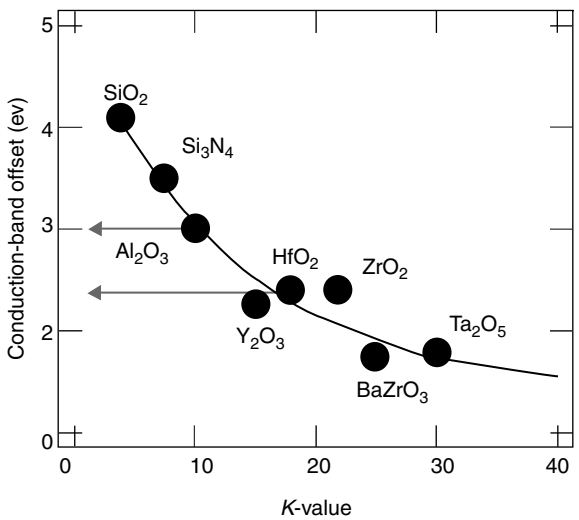


FIGURE 13.42 *K*-value and band gap offsets of high-*k* gate dielectric materials. (Ref. A15). (From Chen, J., 2003 AVS Topical Conference on Atomic Layer deposition, (32) 16-28-1086.)

or silicon oxide, i.e., nanolaminate composite layers has demonstrated to suppress crystallite growth. For hafnium aluminates, the material is maintained in amorphous phase after 900°C, 1 min anneal with [Al] $>$ 50%. For zirconium aluminates, however, a higher aluminum concentration is required to prevent the material from crystallization. Both silicates and aluminates also provide improved interfacial layer control, which is very important to minimize the increases of EOT. Figure 13.43 is a series of TEM images recorded from Hf and Zr aluminates with different Al concentrations. The samples are annealed at 900°C for 60 s The tendency of crystallization is suppressed by increasing Al concentration. The major drawback on the use of silicates and aluminates, however, is lowering film permittivity as a result of lower-*k* SiO_2 and Al_2O_3 incorporation.

Atomic layer deposition have attracted great interests on depositing high-*k* gate dielectric materials. Besides the advantages of depositing smooth, conformal films with precise thickness and composition control, the technology also offers the flexibility on forming silicates and aluminates by using oxygen containing silicon and aluminum precursors, such as silanol and aluminum alkoxide precursors. In addition, silicates and aluminates can also be formed by nanolaminate composite films. An example is shown in Figure 13.44. The materials form miscible amorphous alloy states after annealing.

The precursors applied for ALD high-*k* oxide deposition include halide, alkyl, alkoxide and nitrate compounds. Table 13.9 shows metal precursors and oxidants for high-*k* dielectric deposition reported in the literature.

In most of thin ALD metal oxide deposition, a thin SiO_2 interfacial layer is formed by oxidizing silicon substrates. Figure 13.45a shows ZrO_2 grown on HF-dipped Si substrate using $ZrCl_4$–H_2O chemistry. An interfacial layer of approximately 15–20 Å is observed. In contrast, no interfacial oxide layer is observed

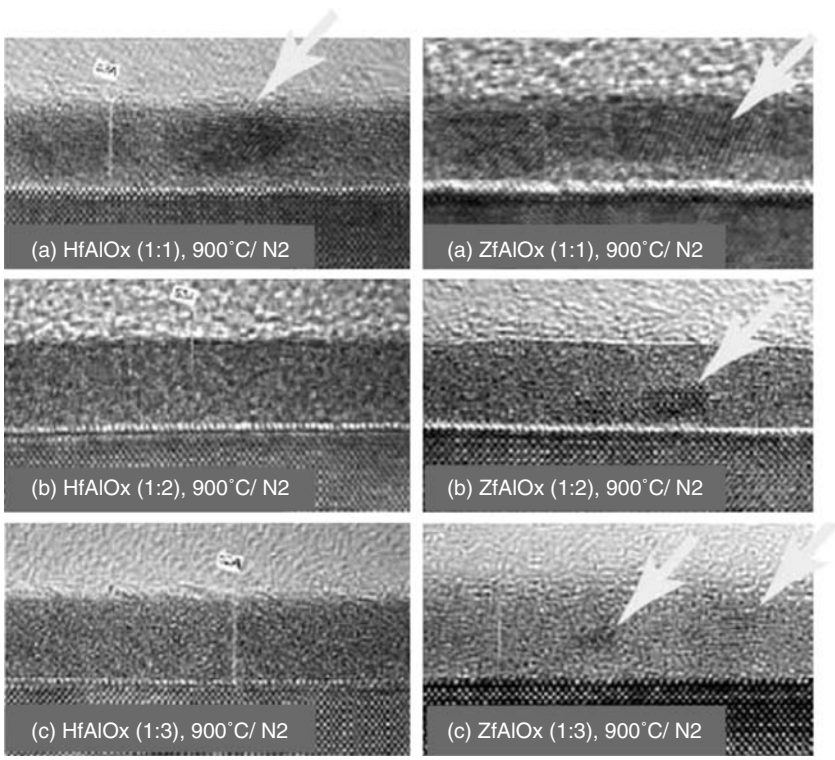


FIGURE 13.43 TEM images recorded from $HfAlO_x$ and $ZrAlO_x$ to show thermal stability after annealed at 900°C for 60 s (Ref. A15). (From Chen, J., 2003 AVS Topical Conference on Atomic Layer deposition, (32) 16-28-1086.)

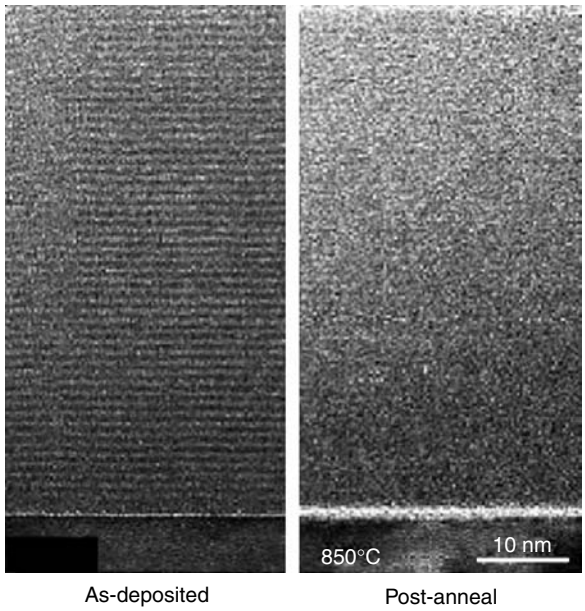


FIGURE 13.44 ZrO_2/Al_2O_3 nanolaminate composite layer deposited by ALD (Ref. A17). (From Zhao, C., *Appl. Phys. Lett.*, 80, 2374, 2002.)

in the growth of Al_2O_3 by using H_2O as the oxidant. The absence of interface oxide layer is expected to be related to the strong reducing nature of TMA which may reduce SiO_2 formed during H_2O exposure. Figure 13.45b is a TEM image recorded from ALD Al_2O_3 grown on HF dipped Si substrates using TMA–H_2O chemistry. No interfacial layer is observed.

13.4.1.8 Polysilicon

In complimentary metal oxide semiconductor (CMOS) technology, doped polysilicon films are mainly used as gate electrodes, capacitor plate electrodes and interconnect due to their high temperature stability (which makes them compatible with further high temperature processing), excellent interface with silicon oxide and chemical purity though their resistivity is much higher than other metals. Other applications for polysilicon films include emitters in advanced bipolar applications, micro-machined micro-electrical mechanical systems structures and load resistors in static memories. As the device

TABLE 13.9 Metal Precursors and Oxidants for High-*k* Dielectric Deposition

Film Type	Metal Precursor	Oxidant
Al_2O_3	$AlCl_3$, $Al(CH_3)_3$, $Al(OC_2H_5)_3$	H_2O, O_2, NO_2, O_3
	$Al(C_2H_5)_3$, $Al(CH_3)_2H$	H_2O_2, N_2O, O_2 plasma
ZrO_2	$ZrCl_4$, $Zr(NMe_2)_4$, $Zr(t\text{-}OC_4H_9)_4$	H_2O, O_2, O_3, H_2O_2, O_2 plasma
	$Zr(OC(CH_3)_3)_4$	
HfO_2	$HfCl_4$, $Hf(NMe_2)_4$, $Hf(t\text{-}OC_4H_9)_4$	H_2O, O_2, O_3, H_2O_2, O_2 plasma
	$Hf(NO_3)_4$	
$ZrSiO_x$	$ZrCl_4$, $Zr(NMe_2)_4$	$(^tBuO)_3SiOH$, $Si(OC_2H_5)_4$
$HfSiO_x$	$HfCl_4$, $Hf(NMe_2)_4$	$(^tBuO)_3SiOH$, $Si(OC_2H_5)_4$
	$HfCl_2(N(SiMe_3)_2)_2$	H_2O
$HfAlO_x$	$HfCl_4$	$Al(OC_2H_5)_3$
	$Hf(NO_3)_4$	$AlCl_3$

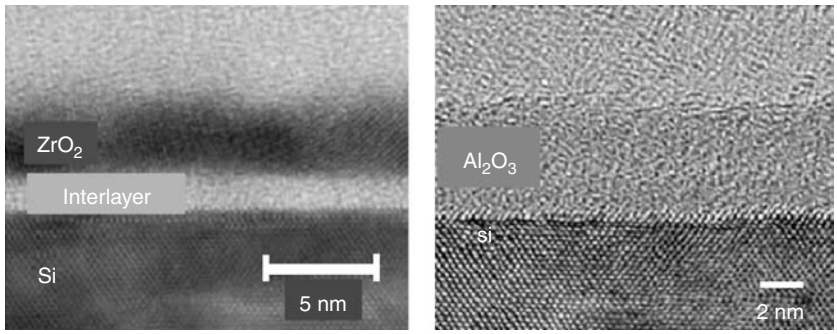


FIGURE 13.45 TEM images recorded from (a) ZrO_2 (Ref. A35). (From Jeon, H., 2003 AVS Topical Conference on Atomic Layer deposition, (82) 2-2290-0387) and (b) Al_2O_3 recorded from HF dipped Si substrates (Ref. A30). (From Jeon, S., 2003 AVS Topical Conference on Atomic Layer deposition, (82) 2-2290-0387.)

geometries shrink, a multi-layer thin film stack called "polycide," (polysilicon + silicide) is used, explained in detail in the section on tungsten silicide.

Polysilicon films can be deposited using hot wall reactors such as vertical and horizontal furnaces and cold wall single-wafer reactors. The process conditions and the structure of the films, of course, depend very much on the type of the reactor used for the deposition of polysilicon films. The details of a single-wafer polysilicon process and the properties of the film obtained are discussed below.

13.4.1.9 Process Conditions

There are many possible source gases for polysilicon deposition—SiH_4, Si_2H_6, SiH_2Cl_2, $SiHCl_3$, $SiCl_4$ etc. In an example of single-wafer polysilicon deposition, SiH_4/H_2 chemistry is used, where SiH_4 is the Si source gas, and H_2 is the carrier gas which also plays a role in the reaction. This is evident from the simplified overall reaction:

$$SiH_4(\text{vapor}) \rightarrow Si(\text{solid}) + 2H_2(\text{gas})$$

For obtaining doped polysilicon films, in situ doping technique is used. In situ doping involves using a dopant source gas such as PH_3, AsH_3 or B_2H_6 along with SiH_4 and H_2. The dopant concentration in the film can be tuned by varying the dopant source bottle concentrations and gas flows.

Typical process conditions for the single-wafer process are as follows:

Temperature	580°C–680°C
Pressure	20–200 Torr
H_2 flow	5–20 slm
SiH_4 flow	200–900 sccm
Dopant flow	10–100 sccm (1% PH_3 in H_2 as example)
Deposition rate	500–1500 Å/min (tunable)

This is quite different from the standard LPCVD process for polysilicon using a furnace. The deposition rates are in the order of 1 Å/min compared to the single wafer process mentioned above. Other differences are highlighted below. Furnaces typically use:

- Pressures between 100 and 300-mTorr
- Temperature between 550 and 630°C
- No carrier gas

The discussion below mainly pertains to the single-wafer, high-pressure poly-Si deposition process. However some of the process trends are very similar at lower pressures and hence is applicable to furnace LPCVD.

13.4.1.10 Effect of Process Parameters

Deposition rate of polysilicon films depends on all the parameters listed above. Poly-Si deposition using the single-wafer technique is surface-reaction controlled and hence temperature is the dominant factor. The dependence of deposition rate on temperature for undoped polysilicon films is given in Figure 13.46. The activation energy (around 660°C) is 2 eV for this process, compared to 1.6 eV obtained for LPCVD poly-Si deposition. To put this in perspective, 1°C change causes 2.2% change in deposition rate. From Figure 13.46, it can be seen from the flattening of the curve that, increasing the pressure lowers the temperature for the transition from surface-reaction limited to mass transport limited mechanism of deposition. It is also evident that increasing the pressure also increases the deposition rate. Figure 13.46 also suggests that higher deposition rates (required for higher throughput) can be obtained with higher temperatures and pressures. However, care must be taken in process optimization as the wrong choice of process parameters can cause gas phase nucleation and particle generation.

Gas flows are obviously important for the polysilicon deposition process. Increasing the SiH_4 flow or decreasing the H_2 flow increases the partial pressure of SiH_4 and thus the deposition rate of polysilicon. In the flow regimes listed above, the relationship between deposition rate and SiH_4 flow is linear as shown in Figure 13.47.

The resistivity of the doped polysilicon films depends both on the dopant concentration and the structure of the films. The main control of dopant concentration is the dopant flow during deposition. Higher dopant gas flow during deposition incorporates more dopant into the polysilicon film and leads to lower resistivity of polysilicon films. In Figure 13.48, the sheet resistance of phosphorous doped films is shown as function of PH_3 (1% in H_2) flow (Temp—660°C, Pressure—80 Torr, SiH_4—500 sccm, H_2—9.5 slm, Thickness—2000 Å. Annealing Condition—850°C, 30 min in N_2/O_2).

It can be seen that the sheet resistance (hence resistivity) of the films decrease with increasing PH_3 flow till a saturation value is obtained. Dopant atoms incorporated in the film may be electrically inactive or active depending on the process conditions. Inactive dopants are typically activated during further thermal processing such as an anneal in a furnace at 850°C for 30 min in N_2/O_2 ambient or in a RTP system at 1050°C for 30 s in N_2/O_2 ambient. This anneal further decreases the resistivity of the polysilicon films. In addition to controlling the resistivity, dopant flow also affects the deposition rate of polysilicon

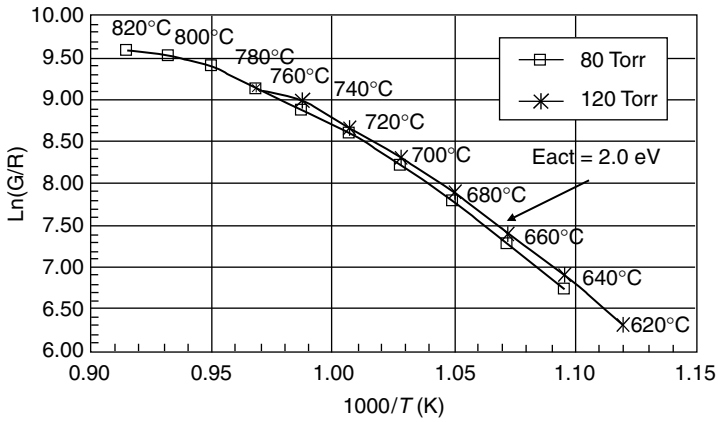


FIGURE 13.46 Deposition rate as a function of temperature.

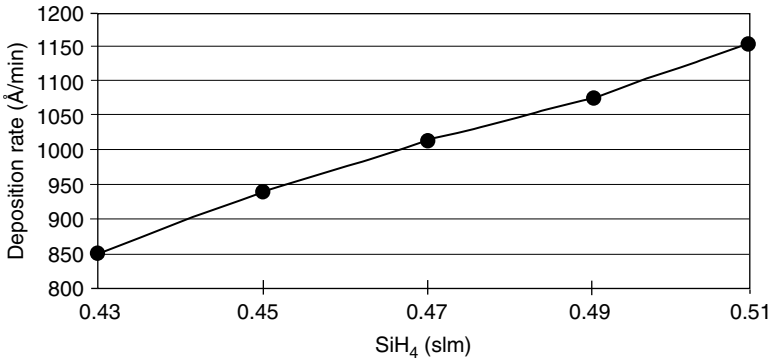


FIGURE 13.47 Deposition rate as a function of SiH_4 flow.

films. Deposition rate decreases with PH_3 and AsH_3 flows and increases with B_2H_6 flow for boron-doped polysilicon films.

Apart from the dopant concentration, the other factor which affects the resistivity of the film is the grain structure. Polysilicon, as the name suggests, has a polycrystalline structure, i.e., it is made up of many small crystals. The size of these crystals, or grains, and their orientation also called texture affects the electrical characteristics of the film.

Depending on the process conditions, sometimes the film can also be deposited as amorphous Si. The competition between surface diffusion of deposited atoms and surface reaction to deposit Si atoms controls whether the film is amorphous or polycrystalline. Faster surface diffusion (low deposition rate, high temperature) favors nucleation and crystallization to form poly-Si films, whereas lower surface diffusion rate (low temperature, high deposition rate) causes amorphous Si films. These films are eventually annealed during thermal processing to give electrically conductive films. The temperature of

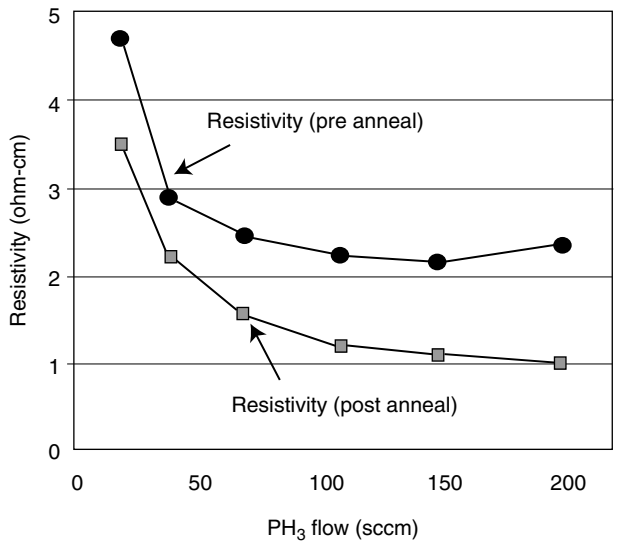


FIGURE 13.48 Resistivity as a function of PH_3 (1% in H_2) flow.

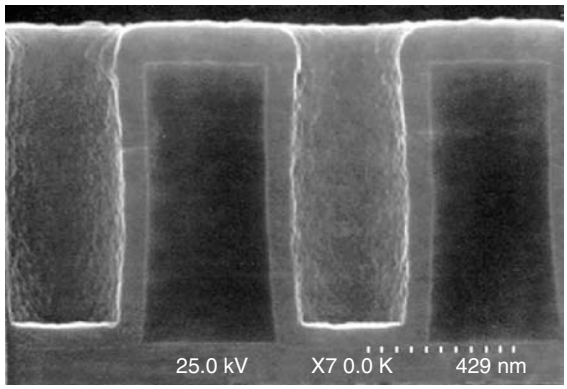


FIGURE 13.49 SEM micrograph showing step coverage >90% for doped polysilicon films.

transition between amorphous and polysilicon films depends on the deposition rate and also the dopant concentration. Dopants typically lower the amorphous-poly Si transition temperature. For the process conditions given in Figure 13.41, the transition temperature for amorphous-poly transition is around 675°C.

Step coverage is another very important parameter as the polysilicon films are used as bitline contacts and capacitor electrodes. Step coverage depends on all the process parameters such as temperature, pressure, etc., Figure 13.49 is a SEM micrograph showing step coverage >90% for a doped polysilicon film obtained using single-wafer deposition.

13.4.2 Conducting CVD Films

13.4.2.1 Tungsten

The tungsten CVD film is known for its excellent step coverage [91]. For <0.8 μm contact or via holes that have aspect ratios greater than two, chipmakers have found it increasingly difficult to use conventional Al sputtering for continuous coating inside the feature and to maintain its electrical performance. The effective via resistance and electromigration resistance have been improved by the introduction of CVD tungsten. The CVD tungsten process has been a key technology enabling multilevel interconnect metallization. There are two approaches of CVD tungsten filling the holes: selective deposition at the bottom of holes by utilizing tungsten hexafluoride's preference to react with metal (aluminum, silicon); and blanket deposition conformally covering all the features simultaneously. During the research stage, there were significant interests in academia in selective tungsten deposition. But eventually the controllability issues of the process in a manufacturing environment became insurmountable. Blanket CVD tungsten fill has become a standard technology for logic devices, and has been adapted to memory contact applications. It has proven to be reliable and extendible to 0.2 μm [92,93] and below. Even in the transition to copper from aluminum, CVD tungsten is still applied at contact level for <90 nm.

13.4.2.1.1 W-CVD Process

Chemical vapor deposition W chemistry has two critical components: SiH_4/WF_6 and H_2/WF_6. The H_2/WF_6 chemistry is an excellent example of process kinetics that switches from a transport limited regime to a surface reaction limited regime. The deposited film properties change along with the regimes, and the chemistry demonstrates an almost perfect hole filling capability and substrate surface-sensitive behavior. Most users begin with SiH_4/WF_6 chemistry to initiate the W deposition (nucleation layer) and

to reduce surface sensitivity; the process then switches to H_2/WF_6 chemistry to achieve the desired film with more repeatable performance.

The H_2/WF_6 process has a very wide operation range:

H_2/WF_6	500/30–5000/500 sccm
Wafer temperature	350°C–450°C
Pressure	5–100 Torr
Typical thickness	1500–10,000 Å

At low temperature and low pressure, the reaction is surface reaction limited and the deposition rate is expressed as:

$$\text{Rate} = k(P_{H_2})^{1/2}\exp(-E_a/kT)$$

The deposition rate is only the function of H_2 partial pressure at constant wafer temperature [94–97]. The formula still holds true even at much higher temperature and pressure if there is sufficient WF_6 supply to the surface. The original transition plot from transport limited regime to surface reaction limited regime can be reformed as shown in Figure 13.50. An example of filling in a high-aspect ratio contact hole is seen in Figure 13.51.

Under conditions in which H_2 and WF_6 are sufficiently diluted, the W deposition rate will be linearly proportional to the supply of WF_6 flow at low flow rates where the supply of gas phase WF_6 is dictating the deposition process [98]. The W deposition rate is much less sensitive to the WF_6 flow at higher flow rates where the adsorbed H_2 reaction with ample WF_6 on the W surface becomes a bottleneck. The transition point will shift to higher WF_6 flow on similar plots at higher temperature, pressure or H_2 flow.

The W film properties such as resistivity, reflectivity, and step coverage follow the same trend. Beginning with the mass transport limited regime and progressing into the surface reaction limited regime, reflectivity at similar as-deposited thickness starts high (smoother film) and progresses to low (rougher film); step coverage starts low and becomes high (close to 100%); stress starts at low tensile and progresses to higher tensile ($1–1.2\times10^{10}$); while fluorine content moves from low to high and resistivity also moves from low to high.

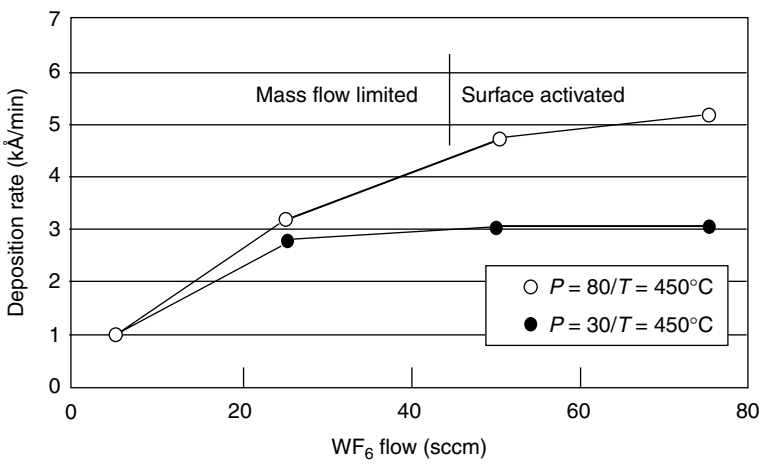


FIGURE 13.50 W deposition rate vs. WF_6 flow at various H_2 flow at constant pressure and temperature.

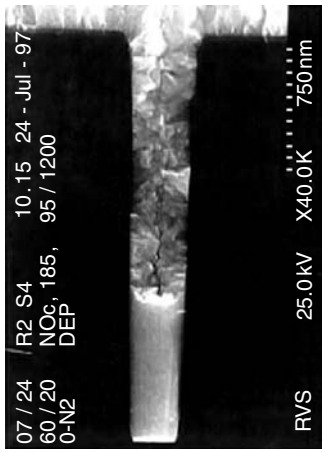


FIGURE 13.51 Example of W filling in high aspect ratio contact hole.

Typical film properties:

Resistivity	8–15 μΩ-cm
Step coverage	50%–100%
Phase	Alpha (BCC)
F content	1×10^{19}–1×10^{20} atoms/cm^3
Stress (~1 μm)	3×10^{9}–1.2×10^{10} dynes/cm^2 tensile

Film properties also depend on thickness. A thinner film has smaller grain, higher stress, higher resistivity and higher reflectivity [99]. At higher deposition temperatures, the deposition process can be driven deeply into the mass transport limited regime and the stress and resistivity can be quite reduced.

The film's thickness dependency is a function of initial state of deposition or "nucleation." Historically, the underlayer of W is TiN which acts as an adhesion promoter between W and SiO_2 [100] and as a reaction barrier between WF_6 and Si or Al. Other materials such as W or TiW alloy have been used, but are less popular. There are many different ways to prepare the TiN layer; these may exhibit quite different surface properties. The SiH_4/WF_6 nucleation is designed to accommodate and minimize the variation of the underlayer.

Chemical vapor deposition-W resistivity in the range of 10–15 μΩ-cm for thick films (>3000 Å), compared with bulk resistivity (5–7 μΩ-cm), tends to be higher; possibly due to smaller grain size and F impurities. One can lower the resistivity by operating deep in the transport limited regime to lower F content, or by manipulating the nucleation/barrier process to enlarge grain size. The addition of B_2H_6 [101] in the nucleation process is an effective means to enlarge W grain size and allows 30%–50% reduction in sheet resistance in very thin (<1000 Å) film applications [102].

In a standard tungsten application for plug fill, the resistivity of W itself does not contribute significantly enough to deteriorate the total via resistance or total line resistance. Some devices take advantage of tungsten's excellent electromigration resistance and/or higher temperature compatibility to use it for longer range interconnect wiring such as high power interconnect, local interconnect, bit line or word line. In those applications, the resistivity of W could negatively impact device performance. By properly mixing the film deposited in two different regimes to optimize effective resistance for filling holes and reaching a desired total film thickness, "tungsten as metal one (WAMO)" has been realized [99]. Or a lower W resistivity, especially in thinner films, can be attractive for those applications.

13.4.2.1.2 Tungsten Plug Formation

The typical process sequence of forming W plugs [103] is after cleaned via/contact opening:

- HF dip on Si surface or Ar sputter clean metal surface;
- Ti deposition to ensure native oxide removal on aluminum or Si;
- TiN deposition as W adhesion promoter and WF_6 attack barrier on Ti or Si;
- W blanket deposition;
- W layer removals by CMP or blanket etch back; and surface cleaning

followed by standard aluminum metallization stack deposition.

During the process sequence, the limitations of material characteristics and process methodology will generate many issues that impact device performance or yield. In one example, where W has poor adhesion on silicon dioxide, it is necessary to prevent direct W deposition on the oxide surface, even at the very edge of wafer. This creates a requirement for an adhesion layer such as TiN and W to have edge exclusion compatibility: if TiN has 1 mm edge exclusion, W deposition has to be excluded more than 2 mm due to stacking tolerance. Similarly, the W deposition chemistry is so aggressive that Ti and Si will be etched away if they are directly exposed to WF_6. This also creates a requirement for TiN deposition to cover all of the underlayer Ti surface: Ti must have less edge exclusion than TiN. Ideally, TiN deposition should have full coverage and cover even the vertical bevel edge portion.

Furthermore, the W removal process has similar wafer bevel compatibility issues: chemical mechanical polishing may not be able to completely clean the bevel of tungsten, which will force tungsten deposition to stay clear of wafer edge bevel. Tungsten etch back does not have this issue of bevel cleaning, but it has relatively low selectivity on silicon dioxide vs. titanium nitride. If the TiN is not full coverage, the wafer edge area will have a trench into silicon dioxide which will impact further wafer processing. In another example, to optimize the final plug performance, the uniformity of tungsten deposition and removal should be matched. If the W removal rate at the wafer edge is much faster than the center on a perfectly uniform tungsten wafer, the plug recess could be very deep at the wafer edge and might affect yield and reliability of the subsequent metallization process.

13.4.2.1.3 Barrier Layers Before CVD-W

Since WF_6 is very reactive with common conductive materials such as silicon [104], aluminum or titanium [105], even small amounts of WF_6 can react with them to form insulating AlF_3 on Al via surfaces, sublimating or insulating TiF_4 on Ti surfaces, or generate SiF_4 gas, which can leave voids (or wormholes) in the Si contact, leading to poor contact or via resistance, high junction leakage and poor structural integrity such as so-called "volcano" [106]. A "good" barrier layer after contact or via opening is therefore required before CVD-W deposition. This barrier should have enough conformality and blockage capability to prevent WF_6 from reaching underlayer Al, Si or Ti. More specifically, at the top corner the barrier should have minimum overhang, good coverage, and no stress cracking; underlayer Ti should also minimize the overhang to ease the barrier process window. At the bottom corner, the barrier should have no cusping and have enough thickness to be an effective barrier. The ideal barrier will be conformal and dense with resistivity low enough to have no impact on device electrical properties; the best candidate so far is CVD TiN.

13.4.2.1.4 W Nucleation Process

The W nucleation process performance can be altered by the barrier material. Among the commonly used TiN adhesion/barrier layer types, such as thermally converted TiN from Ti, reactive sputtered TiN and CVD TiN, each exhibits quite a wide range of material structure, density, stress and conformality. A W nucleation process that is not carefully tuned could result in island-like, non-continuous films which allow potential WF_6 interaction with the underlayer. Thus, the barrier process and the CVD W nucleation process have very tight interdependency with each other so that some minor tuning could be required for device generation-to-generation transitions.

From a chemistry standpoint, the SiH_4/WF_6 nucleation reaction [107] is the same as in the tungsten silicide case (see below). For blanket tungsten nucleation layers, the operational conditions are modified from a standard tungsten silicide process to serve its particular purpose, such as lowering SiH_4 flow to achieve a composition closer to pure W; adding H_2 to blend H_2/WF_6 and SiH_4/WF_6 reactions for better step coverage; and increasing temperature or pressure to match the H_2/WF_6 process conditions enabling one chamber operation. A typical W nucleation process condition is as follows:

SiH_4/WF_6	5:1–1:2
Wafer temperature	350°C–450°C
Pressure	4–80 Torr
Typical thickness	100–500 Å

Since the operational range could fall right into the SiH_4/WF_6 gas phase reaction regime, care should be paid to the hardware design and process conditions such as pressure, concentration, dilution and temperature, to assure minimizing the particle impact and maintaining the desirable surface coverage and nucleation layer properties.

13.4.2.1.5 ALD W as Nucleation Process

The SiH_4/WF_6 nucleation process ran into limitation about 90 nm geometry, due to its insufficient step coverage. ALD W as nucleation layer is introduced. In this section, process characterization of ALD W deposited using WF_6/B_2H_6 chemistry is discussed.

13.4.2.1.5.1 Precursor Exposure Time

A signature characteristic in an ALD system is saturation of the film growth rate (per cycle) under reactant exposures. The film thickness as a function of WF_6 and B_2H_6 exposures is plotted in Figure 13.52. The deposition temperatures are 300, 350, and 400°C, respectively. The films were deposited with a total of 40 cycles of WF_6 and B_2H_6 exposures.

As shown in Figure 13.52, both WF_6 and B_2H_6 exposures reach saturation at less than 0.5 s WF_6 exposure curve follows an ideal self-limiting manner without any further growth after saturation exposure is reached. In contrast, B_2H_6 exposure curves suggest that B_2H_6 thermal decomposition contributes to further film. First of all, the surface reaction of B_2H_6 exposure is completed within 0.5 s with very steep curve. Once surface reaction is completed, B_2H_6 thermal decomposition became dominant leading to further film growth. It should also be noted that reducing exposure time significantly reduces the contribution of B_2H_6 decomposition to film growth at 400°C. B_2H_6 exposure studies suggest that

- Process temperature sensitivity can be reduced by optimizing B2H6 exposure time;
- A lower process temperature $\sim$300°C is preferred for ALD W process.

13.4.2.1.5.2 Temperature Window

Figure 13.53 shows tungsten growth rates at various temperatures. For temperatures below 250°C, slow growth rates accompanying with poor thickness uniformity indicates an incomplete reaction. At medium temperature range between 250 and 350°C, the growth rate is approximately 2.3–2.8 Å/cycle, which corresponds to the thickness of a tungsten monolayer. Good Rs and thickness uniformity ($\leq$1.5%, 3 mm edge exclusion) is obtained. A minor temperature dependency is attributed to minor B_2H_6 thermal decomposition in the corresponding temperatures. A minimization of B_2H_6 exposure significantly reduced the temperature dependency with no impact on film properties. Above 350°C, high growth rates associated with poor thickness uniformity as a result of significant B_2H_6 thermal decomposition are observed. A temperature window between 250 and 350°C is suitable for ALD W deposition.

Figure 13.54 presents the proposed reaction pathways in the tungsten film deposition. A single layer tungsten film is deposited during the WF_6 exposure, with further film growth inhibited by surface fluorine coverage. B_2H_6 reacts with the surface fluorine to form volatile byproducts (HF, BF_3, etc.) and

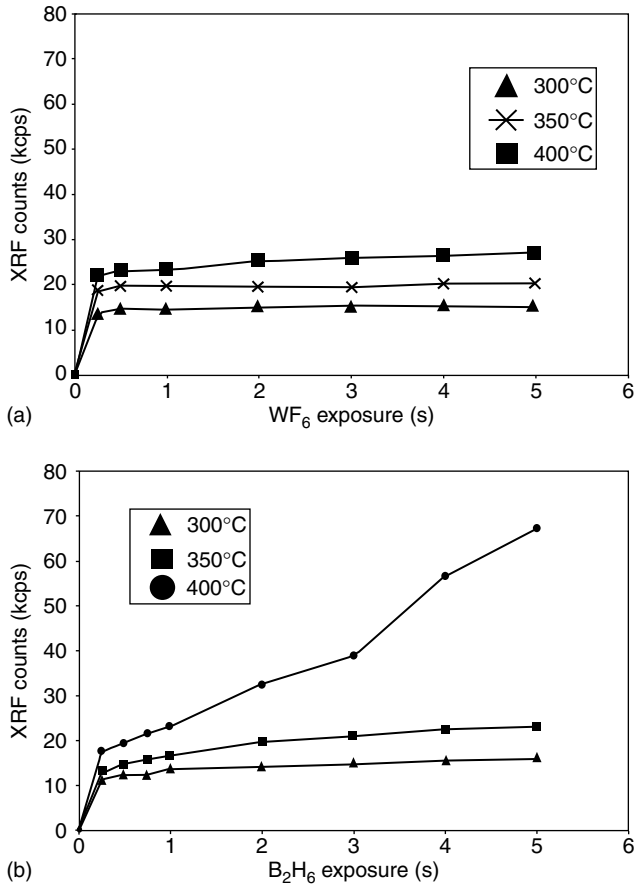


FIGURE 13.52 Film deposition rate vs. (a) WF_6 and (b) B_2H_6 exposure time in ALD W process.

regenerates the surface for the next WF_6 exposures. Excess boron-containing species may remain on the surface and contribute to the observed increases in growth rate with increasing B_2H_6 exposure in Figure 13.54.

It is found that fluorine concentration is increased if significant B_2H_6 decomposition is involved during film deposition. Table 13.10 summarizes fluorine concentration (atoms/cm^3) measured by secondary ion mass spectroscopy from ALD W deposited at different temperatures with 0.25 and 5 s of B_2H_6 exposures, respectively.

Figure 13.55 shows the tungsten film thickness as a function of ALD cycle count for a typical deposition condition with 0.25 s exposures for WF_6 and B_2H_6. As expected for ALD processes, the film thickness increases linearly with deposition cycle count. The film growth rate is 2.5–3.0 Å/cycle, with a negligible film incubation period, consistent with the lattice spacing of tungsten.

The combination of decreasing diameter and increasing aspect ratio poses challenges for conventional tungsten plug fill. Tungsten deposition technology is currently based on reduction of WF_6—by silane (SiH_4) for the nucleation and by hydrogen (H_2) for the bulk deposition stage. The key contributor to tungsten plug fill capability is the step coverage of the nucleation layer, which for the current CVD-based technology can be significantly less than 50%, imposing significant challenges to complete plug fill. Compared to conventional CVD processes, ALD provides highly conformal, ultra-thin films with very

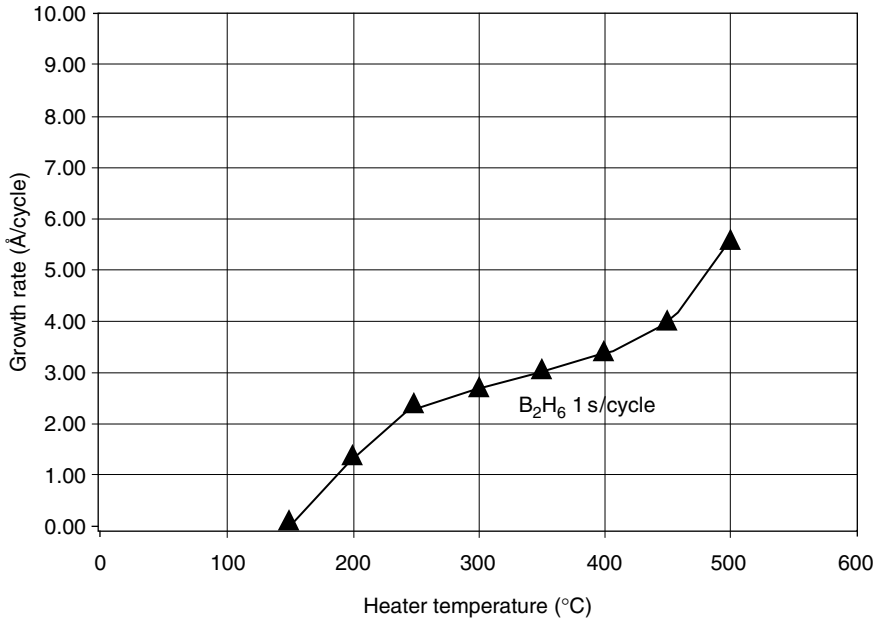


FIGURE 13.53 Growth rate/cycle as a function of deposition temperature. The B_2H_6 exposure time is fixed at 1 s/cycle.

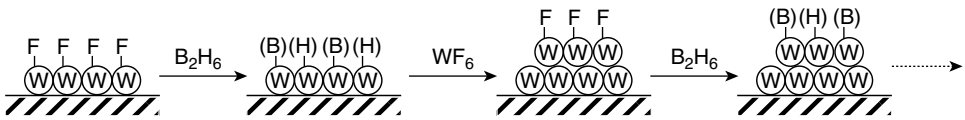


FIGURE 13.54 Proposed reaction mechanism of W film growth in WF_6/B_2H_6 chemistry.

accurate thickness control. These unique features enables ALD W as a nucleation layer for tungsten plug fill of silicon contacts in advanced semiconductor devices.

X-ray diffraction spectrum recorded from an ALD W sample deposited using WF_6/B_2H_6 chemistry is shown in Figure 13.56a. The film is of amorphous structure and is stable even after 800°C annealing (Figure 13.56b). A plane view TEM micrograph with an electron diffraction pattern recorded from a similar sample is shown in Figure 13.56c. The result confirmed that the film is amorphous in nature.

The fluorine level in the ALD film is $<2\times10^{18}$ atoms/cm^3 as measured by SIMS. The film resistivity is approximately 150 μΩ-cm. The ALD tungsten surface is smooth with a roughness of 6 Å (rms) for 1000 Å film. This indicates a uniform layer-by-layer film growth over the substrate. The film stress is 5×10^9 dynes/cm^2 as measured from 250 Å ALD W sample grown on top of metal organic chemical vapor deposition (MOCVD) TiN.

TABLE 13.10 Fluorine Concentration Comparison from ALD W Film Deposited at Different Process Conditions

	300°C	350°C	400°C
0.25 s	1.00×10^{18}	1.00×10^{18}	2.00×10^{18}
5 s	3.00×10^{18}	5.00×10^{19}	5.00×10^{20}

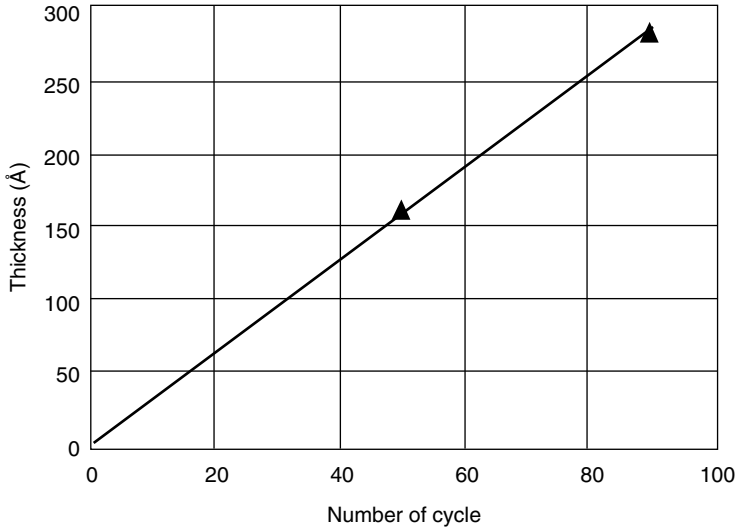


FIGURE 13.55 ALD W film thickness as a function of deposition cycles.

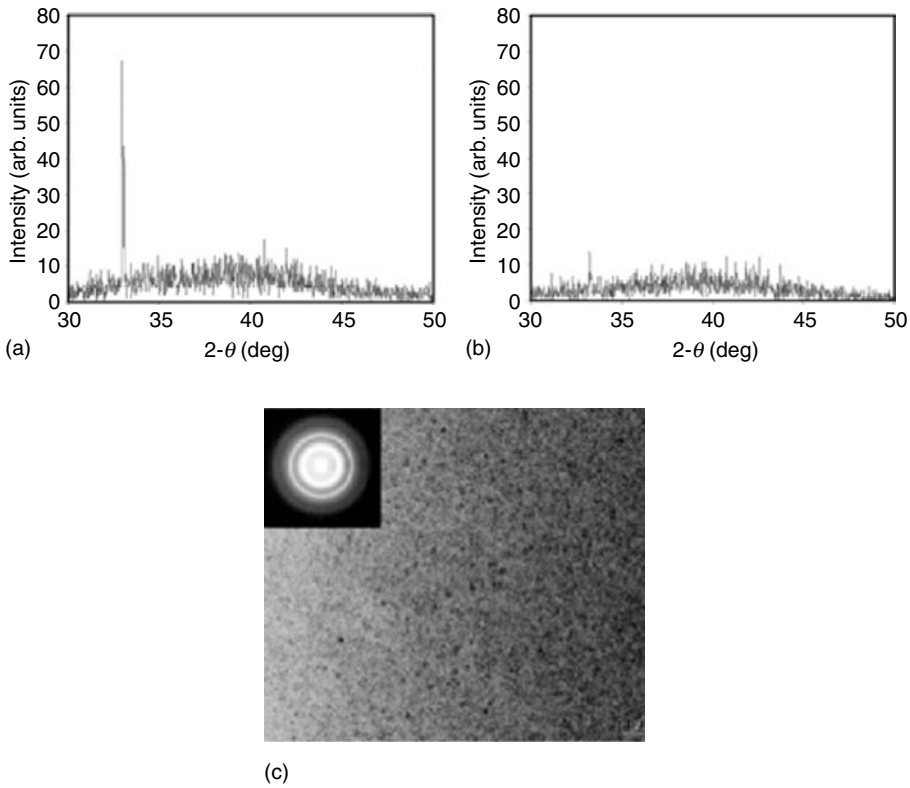


FIGURE 13.56 X-ray diffraction tungsten ALD film (a) as-deposit (b) after thermal annealing at 800 °C and (c) Plane-view TEM of tungsten ALD film.

Atomic layer deposition W deposited using SiH_4 as the reducing agent has also been examined. At 350°C the WF_6/SiH_4 process exhibits self-limiting behavior for exposure times on the order of 0.25 s The thickness of the film increases linearly with number of exposure cycles with a growth rate of 5.6 Å/cycle. The higher deposition rate per cycle with SiH_4 reduction may be a result of parallel reaction pathways: (i) surface reduction of the WF_6 by SiH_4 and (ii) incorporation of SiH_4 into the deposited tungsten film to form WSi_x.

The WF_6/SiH_4 nucleation layer is observed to be crystalline with a grain size on the order of 80 Å and a surface roughness of 15–20 Å for a 1000 Å thick film. The film has a resistivity of ~160 μΩ-cm. The fluorine level in the bulk of WF_6/SiH_4 ALD film is 7×10^{20} atoms/cm^3. The film stress for a 250 Å film is 2.0×10^{10} dynes/cm^2. Comparing the properties and performance of the WF_6/SiH_4 chemistry with the WF_6/B_2H_6 chemistry, initiation of the film growth from the WF_6/SiH_4 chemistry shows a higher sensitivity to the substrate. The WF_6/SiH_4 nucleation layer formed has a higher fluorine content and higher stress than the WF_6/B_2H_6 film.

Stack film properties of various nucleation layers for tungsten CVD, a comparison of bulk deposition film stack film properties for the two tungsten ALD nucleation processes and a conventional CVD tungsten nucleation process (WF_6+SiH_4 reduction) are tabulated in Table 13.11. With the same thickness for bulk film deposition, all three nucleation processes yield similar resistivity. The integrated film stress and fluorine level at the tungsten/TiN interface are lowest for the WF_6/B_2H_6 ALD nucleation. With a 100 Å WF_6/B_2H_6 ALD nucleation layer, the F level at the interface after bulk deposition is 4×10^{20} atoms/cm^3. The fluorine levels are 1×10^{21} atoms/cm^3 at the interface with SiH_4 ALD and conventional tungsten nucleation layers, respectively.

13.4.2.1.6 Application of Tungsten ALD Film as Nucleation Layer for WCVD

Tungsten CVD followed by tungsten CMP is currently used to form contacts to silicon in multi-level interconnect fabrication. One disadvantage of conventional tungsten CVD process is the limited step coverage (<50%) of the nucleation layer for small-diameter, high-aspect ratio contacts/vias. Overhang created by the CVD nucleation process can lead to void formation after bulk tungsten deposition and creates seams after tungsten CMP, a concern in subsequent process steps. In order to achieve a complete contact/via fill, the nucleation layer should be continuous with excellent step coverage. Figure 13.57a shows the conformality of the WF_6/B_2H_6 ALD film deposited on an oxide patterned substrate with a via diameter of 0.25 μm and an aspect ratio of 6:1. The film step coverage is 100%, and no overhang or thinning at the top and bottom corner of the contact is observed.

The excellent plug-fill capability of ALD nucleation integrated with CVD tungsten was demonstrated by TEM images collected after CMP. The pattern shown in Figure 13.57b has a plug size of 0.20 μm and an aspect ratio of ~7–8:1 and Figure 13.57c has a via size of 0.2 μm with an aspect ratio of 30–50:1. The top-view SEM micrograph in Figure 13.58 shows a significant reduction of post CMP dimple size of tungsten plugs with a tungsten ALD nucleation layer. It is expected that tungsten nucleation layer deposition by ALD enables a complete tungsten bulk fill in the high-aspect-ratio features which lead to the reduction of post CMP dimple sizes.

TABLE 13.11 Comparison of the Integrated Film Properties for a 3500 Å Film with SiH_4 CVD, B_2H_6 ALD and SiH_4 ALD Nucleation

	SiH_4 CVD Nucleation	B_2H_6 ALD Nucleation	SiH_4 ALD Nucleation
Film stack resistivity (3500 A)	11 μΩ-cm	10.7 μΩ-cm	11 μΩ-cm
Film stack stress (3500 A)	1.6×10^{10} dynes/cm^2	1.2×10^{10} dynes/cm^2	1.6×10^{10} dynes/cm^2
Film stack interface fluorine content	1×10^{21} atoms/cm^3	4×10^{20} atoms/cm^3	1×10^{21} atoms/cm^3
Film stack bulk fluorine content	2×10^{20} atoms/cm^3	1×10^{20} atoms/cm^3	2×10^{20} atoms/cm^3

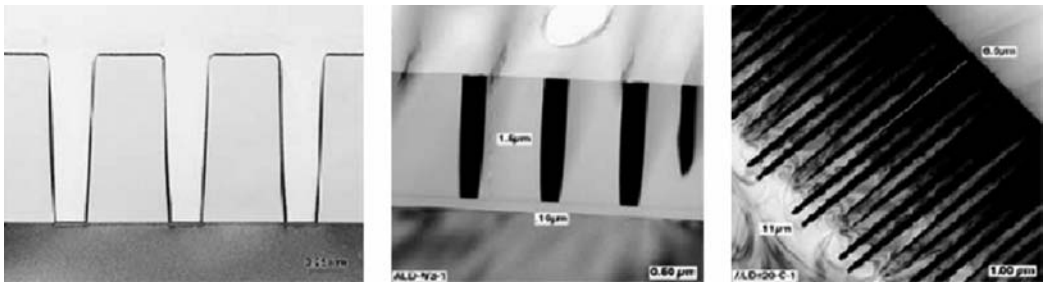


FIGURE 13.57 (a) TEM cross-section of tungsten ALD film on patterned oxide wafer. The full process integration of tungsten ALD with standard WF_6/H_2 via fill process for (b) 0.20 μm×1.5 μm plug and (c) 0.2 μm×7 μm via patterns.

13.4.2.2 Tungsten Silicides

Tungsten silicide is used in combination with polysilicon in so-called "polycide" structures in place of standard doped polysilicon. Its advantage over single-layer doped polysilicon is lower sheet resistance while maintaining polysilicon integration characteristics such as for polysilicon self-aligned gates. The most common applications are for the gate electrode and bit line in memory devices [108]. Tungsten silicide has similar oxidation resistance and wet chemical resistance (HF, RCA clean) as polysilicon and is thermally compatible with polysilicon. One of the reasons tungsten silicide stands out among the various silicides is its manufacturability by CVD.

There are two standard chemistries to deposit tungsten silicide. Monosilane (SiH_4) with tungsten hexafluoride (WF_6) is the conventional one [109], while dichlorosilane (SiH_2Cl_2) with WF_6 is increasingly used to meet the requirements of smaller geometries (<0.5 μm).

13.4.2.2.1 Monosilane/Tungsten Hexafluoride

The SiH_4/WF_6 reaction is thermodynamically extremely favored. The reaction is very exothermic and can be very violent in the gas phase if the mixing between gases is not handled properly. To control the reaction, SiH_4 and WF_6 should be introduced into the reactor separately with mixing taking place at low

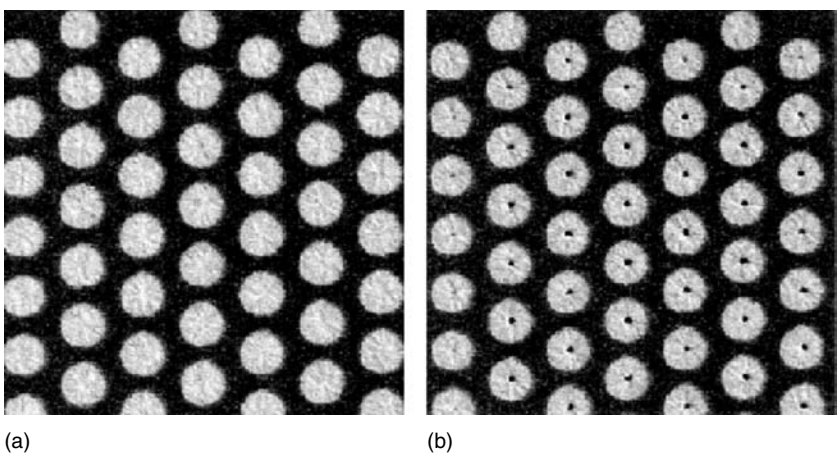


FIGURE 13.58 Top-view SEM of (a) tungsten ALD nucleation and (b) standard tungsten nucleation.

pressure and low temperature right before introducing the gases on the wafer. A cold-wall reactor with the ability to deliver gases individually to each wafer is desired. Since both reactants are very sensitive to moisture and atmospheric leaks, the system should have loadlocks and leak-check capability to minimize the potential problems.

Typical reaction conditions are as follows:

SiH_4/WF_6	500/5–1000/10 sccm
Heater temperature	350°C–400°C
Wafer temperature	300°C–350°C
Pressure	0.5–0.8 Torr
Typical thickness	2000–5000 Å

The SiH_4/WF_6 ratio is the key factor in controlling the deposited silicide composition; normally the ratio is high enough to assure an as-deposited Si/W ratio >2.6. The deposition rate in this regime is linearly proportional to the WF_6 flow and has almost no sensitivity to temperature [110]. The reaction takes place in the supply limited regime and the step coverage is limited by the incident angle of incoming gas species.

Typical film properties are:

Resistivity (as deposited)	800–950 μΩ-cm
Resistivity after 900°C anneal	50–70 μΩ-cm
Si/W (as deposited)	2.6–2.8
Si/W after anneal on Si	2.2
Stress (as deposited)	$6–8\times10^9$ dynes/cm^2 tensile
Stress after anneal	$1.0–1.2\times10^{10}$ dynes/cm^2 tensile
F content	5×10^{20} atom/cm^3
Phase (as deposited)	Amorphous
Phase after anneal	Tetragonal
Step coverage	30% at 1:1 aspect ratio

The target Si/W ratio depends on the integration sequence. If there are many cycles of oxidation after silicide deposition, the Si content could be consumed by the oxidation process and reduce the final Si/W ratio to less than the thermodynamical stable phase composition of 2.2. The polysilicon under the silicide can act as an extra silicon source at oxidation temperature, but the polysilicon–silicide interface could have residues such as oxide to block the diffusion locally. This in turn would cause voids and loss of adhesion at the tungsten silicide–polysilicon interface. Too-high silicon content will reduce the lower-resistivity advantage of tungsten silicide.

The as-deposited tungsten silicide is amorphous [111]. During annealing, tungsten silicide goes through a series of phase transformations typically present in the stress to temperature hysteresis curve as shown in Figure 13.59.

From the amorphous state, the hexagonal phase starts to form around 450°C. The amorphous-to-hexagonal phase transformation has very large volume shrinkage and large tensile stress build up (up to 1×10^{10} dynes/cm^2) and can cause film cracking or discontinuity around step corners [112]. This can degrade the polycide runner over a gate array area to suffer higher line resistance. The hexagonal phase will remain stable until about 650°C, where the tetragonal phase transformation begins. The hexagonal–tetragonal phase transformation has less of a volume change and does not generate as much damage as the previous one.

Fluorine concentration in the as-deposited film is quite high in percentage [112–115]. The fluorine can potentially migrate to the polysilicon/oxide gate interface after high-temperature annealing and increasing the effective gate oxide thickness. The fluorine effect becomes more intolerable when the gate oxide reaches <100 Å and annealing temperature is >850°C. This effect drives the need to reduce the tungsten silicides fluorine content and encourages the use of dichlorosilane chemistry [112].

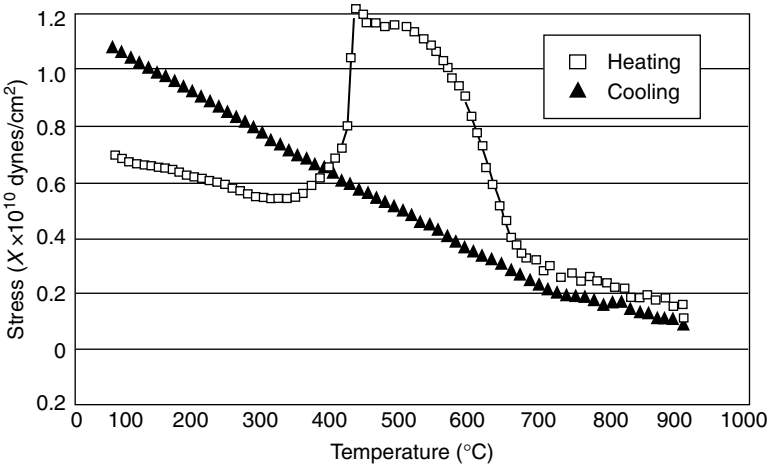


FIGURE 13.59 Monosilane tungsten silicide film stress vs. temperature curve.

13.4.2.2.2 Dichlorosilane/Tungsten Hexafluoride

SiH_2Cl_2/WF_6 is not a simple reaction kinetically [106–108]. No interaction occurs at $<450°C$ without external assistance. To start the reaction requires striking plasma, or much higher temperature, or/and introduction of some ignition agents such as SiH_4/WF_6; the latter is perhaps the most popular method. The nucleation process can be very difficult and is sensitive to the reactor design. Once the process is initiated properly, the deposition conditions are similar to the SiH_4/WF_6 reaction. Typical deposition process parameters are:

SiH_2Cl_2/WF_6	200/3–500/5 sccm
Heater temperature	500°C–600°C
Wafer temperature	450°C–550°C
Pressure	0.5–1.2 Torr
Typical thickness	500–2500 Å

Process kinetics such as deposition rate are sensitive to WF_6 flow and need a high SiH_2Cl_2/WF_6 ratio to achieve the desired Si/W ratio. The SiH_2Cl_2/WF_6 process has additional sensitivity to process temperature and pressure compared with SiH_4/WF_6 chemistry; higher temperature and pressure drive up the deposition rate and Si/W ratio. The typical process is not clearly placed in either the surface reaction limited regime or mass transport limited regime but has characteristics of both types. It only indicates that the reaction pathways are very complicated. The resulting step coverage is much better than SiH_4/WF_6 chemistry, but is not close to 100% (Figure 13.60).

Typical film properties are:

Resistivity as deposited	750–1200 μΩ-cm
Resistivity after anneal	80–120 μΩ-cm
Si/W as deposited	2.5–3.0
Si/W after anneal on Si	2.2
Stress as deposited	$0.8–1\times10^{10}$ dynes/cm^2 tensile
Stress after anneal	1.2×10^{10} dynes/cm^2 tensile
F content	1×10^{17} atoms/cm^3
Phase as deposited	Hexagonal
Phase after anneal	Tetragonal
Step coverage	30% at 3:1 aspect ratio

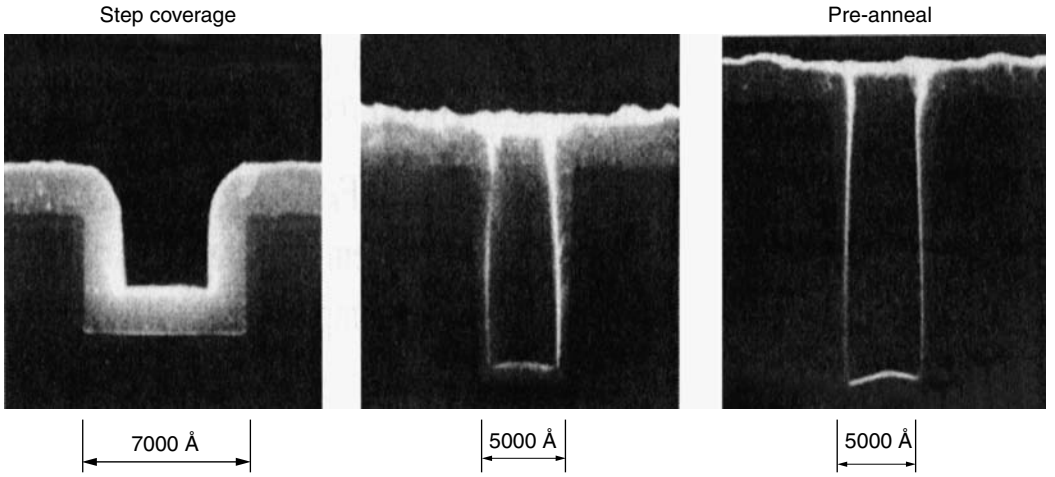


FIGURE 13.60 Step coverage of dichlorosilane tungsten silicide film.

The requirement for the Si/W ratio is similar to the case of SiH_4/WF_6 for the same application. The post-annealed resistivity can be a little higher than SiH_4/WF_6 due to smaller grain size and less chemical driving force to encourage grain growth. The as-deposited film is hexagonal in phase since the temperature is above 450°C and the detrimental amorphous–hexagonal phase transformation can be prevented as shown in Figure 13.61.

The step corner cracking seems to be completely eliminated by better step coverage and better material behavior [119]. Fluorine content is much lower due to the inherent chemistry pathway change, and $<1\times10^{18}$ atoms/cm^3 seems to be well within the lower bound concentration affecting thin gate oxides [112,120–122]. The adhesion also is somewhat improved. Overall, the dichlorosilane chemistry improves the tungsten silicide film properties to be suitable for more aggressive device geometries. However, the trade-offs are higher process temperature and greater process sensitivity.

13.4.2.3 Titanium

Titanium is widely used as a salicidation material to form titanium silicide on the gate and source/drain area, specifically in logic applications for its low contact resistance and low bulk resistance.

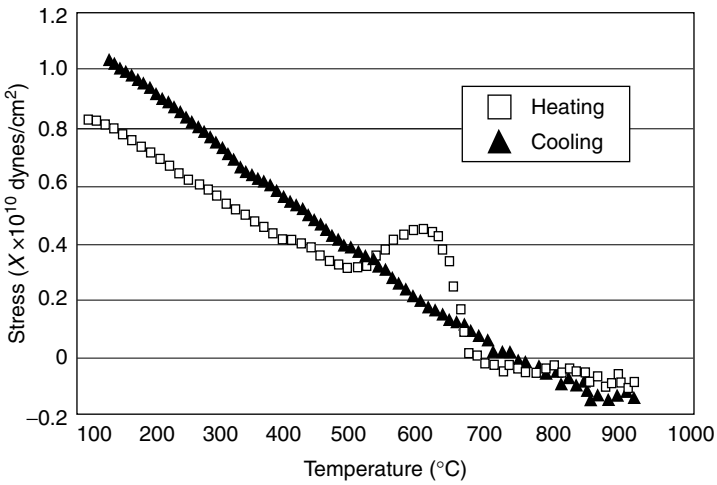


FIGURE 13.61 Dichlorosilane tungsten silicide stress vs. temperature curve.

Plasma-enhanced chemical vapor deposition titanium does not replace the traditional PVD-deposited titanium, but is finding a niche application: forming silicide at the bottom of very high aspect ratio contacts in memory devices [123] where PVD method has difficult time to pile enough material without excessive build-up on the surface.

The PECVD Ti process uses a "brutal force" approach to break up $TiCl_4$ in a H_2 plasma environment and at relatively high substrate temperatures to form titanium silicide in situ [124–130]. Due to the reactive nature of the chemistry and its high temperature requirements, PECVD Ti cannot be used at the aluminum via. At the contact level, specifically in DRAMs, very high aspect ratio contacts at smaller geometries are dictated by large capacitor structures. Since the speed in DRAM devices may not be as critical as logic devices, silicide by PECVD Ti can be used at the bottom of the contact hole for excellent contact resistance improvement.

The detailed process mechanism is not well understood. Highly diluted $TiCl_4$ in H_2 plasma can produce metallic Ti. At more than 500°C wafer temperature titanium silicide will form on the silicon substrate. Under typical process conditions, adding more RF power to the process does not generate more Ti silicide on a silicon substrate, although it does deposit more Ti on an oxide substrate. The Ti silicide formation seems more sensitive to temperature and time, probably limited by Si diffusion in the silicide formation process. At the bottom of a contact, the Ti silicide can be more than twice as thick as Ti on the top oxide surface (Figure 13.62). This indicates that the silicide formation precursors are different from Ti deposition. Some researchers have suggested that a more populated $TiCl_3$ unsaturated Ti halide which would not deposit Ti on oxide surface, in addition to Ti, could react with Si on the substrate surface to form Ti silicide at lower than normal Ti–Si silicidation temperatures [131].

$$TiCl_x + (2 + 4/x)Si \rightarrow TiSi_2 + 4/xSiCl_4$$

$TiCl_4$ itself could selectively react with Si at temperatures >700°C. The Ti subchlorides generated in H_2 plasma could well be the right precursor since they are more abundant than Ti alone and their lifetime could be long enough to diffuse into the contact bottom to complete the reaction.

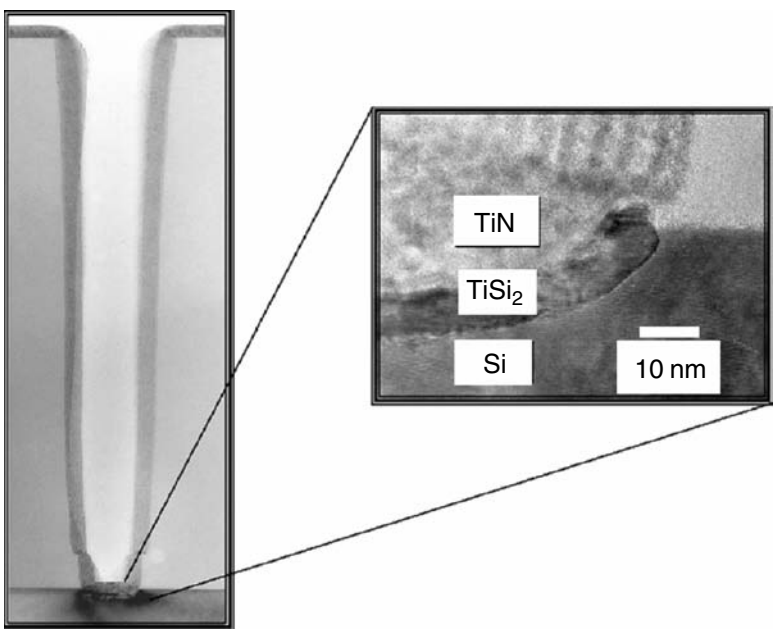


FIGURE 13.62 Cross-section TEM of the bottom of a 0.3 μm×2.3 μm contact showing the uniform silicidation and no elimination of sidewall silicide creep.

The PECVD Ti process creates a harsh environment for process hardware, involving high temperature, halide and plasma all at the same time. Material selection for chamber construction, heater materials and chamber cleaning are continuous challenges for this process.

13.4.2.4 Titanium Nitride

Titanium nitride (TiN) is used in numerous applications:

- As a cladding layer in Al metallization to enhance interconnect wiring electromigration resistance [132];
- As a CVD-W adhesion layer over oxide and barrier against the interaction of WF_6 with Al and Si [133];
- As local interconnect where Al metallization cannot bear the temperature [134];
- As a plate electrode for Ta_2O_5 capacitors [136]. As the node and plate electrodes for MIM capacitor where the insulator is either ALD Al_2O_3 or HfO_2/Al_2O_3 laminate.

Physical vapor deposition is the most common deposition method in most of the applications because of its simplicity. The step coverage requirement for high aspect ratio applications such as deep contacts ($>6{:}1$) for DRAMs, unlanded vias ($>4{:}1$) for logic, and tall cylindrical DRAM capacitors, drive the development of CVD TiN.

Many chemistries have been reported in the literature for CVD TiN: $TiCl_4/NH_3$ [136–140] with or without plasma assist, TDMAT/NH_3 [141–143], and [TDEAT]/NH_3 [144]. A brief comparison between them is shown in Table 13.12.

There is no universal solution for all the potential CVD TiN applications as shown in the table. $TiCl_4/NH_3$ has excellent performance for deep contact applications, but falls far short of multi-level metal applications where temperature must be kept below 450°C. Metal-organic compounds thermal decomposition can achieve low temperature, good comformality, but as-deposited resistivity is not attractive. Metal-organic with ammonia improved the film resistivity at low temperature, but the fast gas phase reactions limit its step coverage and particle performance. Plasma treatment of N_2/H_2 on thermally decomposited TDMAT at temperatures compatible with aluminum metallization can achieve sufficient film properties such as lower resistivity, good step coverage, and improved film microstructure, by combining the advantages of excellent step coverage of metal-organic thermal decomposition and plasma purification of as-deposited TiN. Again, the plasma treatment method suffers in its practical applications only for thin films (<200 Å) due to its long process time.

TABLE 13.12 Comparison of TiN Deposition Chemistries

Chemistry	Comments
$TiCl_4/NH_3$ thermal	Temperature >450°C, not compatible with Al
	Cl content not compatible with Al
	Adduct formation on cold surface
	Excellent step coverage and film properties
$TiCl_4/NH_3$ plasma	Cl content not compatible with Al
	Poor step coverage
TDMAT thermal	Carbon content high
	Resistivity high when exposed to ambient
	Excellent step coverage
	Poor film properties
TDMAT/NH_3	Gas phase reaction particle concern
	Trade off between film properties and step coverage
TDEAT/NH_3	Gas phase reaction particle concern
	Trade off between film properties and step coverage

13.4.2.4.1 $TiCl_4/NH_3$

One distinguishing feature in the interaction of halide compounds with NH_3 is the formation of halide-$(NH_3)_x$ adducts [145]. These adducts will condense inside the reactor, in the foreline, in the pump and in the exhaust line—i.e., wherever the surface temperature is cold. The condensates cannot only generate particulates but can block the pump line to significantly restrict pumping speed and erode the pump itself beyond its design tolerance. The handling of adducts becomes a major problem in dealing with this class of reaction.

The adduct complex will sublimate at moderate temperatures (typically 100°C–200°C) and low pressure. One can control the location of condensation by carefully managing the temperature of the reactor wall and foreline, while placing a water-cooled cold trap in the line to concentrate the condensate in one place. The cold trap can be maintained by simply washing it in water since the adduct is typically highly soluble in water.

The reaction rate is controlled by the $TiCl_4$ and NH_3 partial pressure [136–138] in addition to temperature; the higher the pressure, the higher the deposition rate. Step coverage is affected by the ratio of $TiCl_4$:NH_3, flow, and pressure; surprisingly, there is not a strong function of temperature. The ratio of $TiCl_4$:NH_3 seems to dominate film properties control; more NH_3 equates to less Cl residues in the film, poor stepcoverage and lower resistivity. Typical process conditions are as follows:

$TiCl_4/NH_3$	1:5–1:10
Pressure	1–20 Torr
Wafer temperature	500°C–650°C
Typical thickness	200–600 Å

Most applications prefer the TiN process to operate at the high end of its temperature range because of lower Cl content, lower resistivity, equivalent step coverage and not much concern for temperature at the contact level. The conventional mass transport to surface reaction limited regime transition seems not to be clearly defined in this case. The reasons probably are the low apparent activation energy of thermal decomposition of the halide–NH_3 complex, which cannot be easily distinguishable from the activation energy of diffusion; and/or multiple temperature dependent, interchangeable forms of halide–NH_3 complexes that artificially raise the activation energy of diffusion.

The resulting TiN film has the following characteristics:

Resistivity	150–500 μΩ-cm
Cl content	1.5–5 at.%
Phase	Cubic structure
Stress	$1–2\times10^{10}$ dynes/cm^2 tensile
Stepcoverage	$>$95% at 3:1 aspect ratio

One concern is the high tensile stress which limits the maximum thickness (typically <1000 Å) without cracking. This property has limited the application of the CVD TiN to thin barriers or local interconnects.

13.4.2.4.2 Metal-Organic TiN Films

The metal-organic compounds used as source gases (TDMAT or TDEAT) for CVD TiN normally are liquids at room temperature [146]. Two methods are widely used in the industry to accurately control the supply rate. One involves introducing a carrier gas through the liquid in an ampoule (bubbler), and the other is injecting metered liquid onto a hot surface to convert the liquid into its gaseous phase (liquid injection).

The bubbling method is simple, but the precursor supply rate is not linear with carrier flow and it is not easy to calibrate the quantity. This method is sensitive to the operation pressure, line length, ampoule liquid level (especially at high end of precursor usage), and liquid temperature; a relatively low amount of liquid can be delivered. The liquid injection method avoids most of the bubbler's problems,

but the completeness of vaporization depends on the vaporizer design. A hotter vaporizer can provide more complete vaporization, but is limited by the compound's stability at higher temperatures.

13.4.2.4.3 Thermal Decomposition and Plasma Treatment Cycle

The thermal decomposition of TDMAT or TDEAT is similar to most thermal decomposition reactions. There is a clear transition from mass transport to surface reaction limited regimes. The film properties and step coverage do follow the conventional regime transition. A typical operating condition is close to the transition point to benefit or trade-off from step coverage and low resistivity. However, the thermally decomposed film may still have a significant amount of hydrocarbon attached to the nitrogen. The resulting films have resistivity around 2000 μΩ-cm and can increase to more than 20,000 μΩ-cm once exposed to ambient and oxidized [147,148]. To reduce the carbon content and resistivity, a post-deposition plasma treatment in H_2 and N_2 can effectively reduce impurities and stabilize the film properties. The resistivity of the plasma-treated film can achieve 200 μΩ-cm and carbon content $<5\%$, while preserving the step coverage [149,150].

The N_2/H_2 plasma generates NH active species which impinge on or diffuse into and then replace amino groups in the deposited film. The amino group may diffuse out as a whole or be broken done to more basic groups to diffuse out of the film. Nitrogen isotope tests of the plasma-treated film have confirmed that the nitrogen content in the resulted film is indeed from N_2/H_2 gas phase instead of the metal-organic precursor [151]. The effective depth of treatment is typically about 100 Å and multiple cycles of deposition and treatment are required for thicker films. Fortunately, most TiN applications are for barriers, whose thickness requires only one or two treatment cycles.

One unique characteristic of the deposited/treated film is its structure control. The as-deposited thermally decomposed film is a low-density, amorphous phase of $TiN_xC_yH_z$. The plasma treatment densifies the film by removing bulky hydrocarbon groups, and its energetic ion bombardment compacts the film to become a nanocrystalized matrix, as shown in Figure 13.63. By controlling the thermal

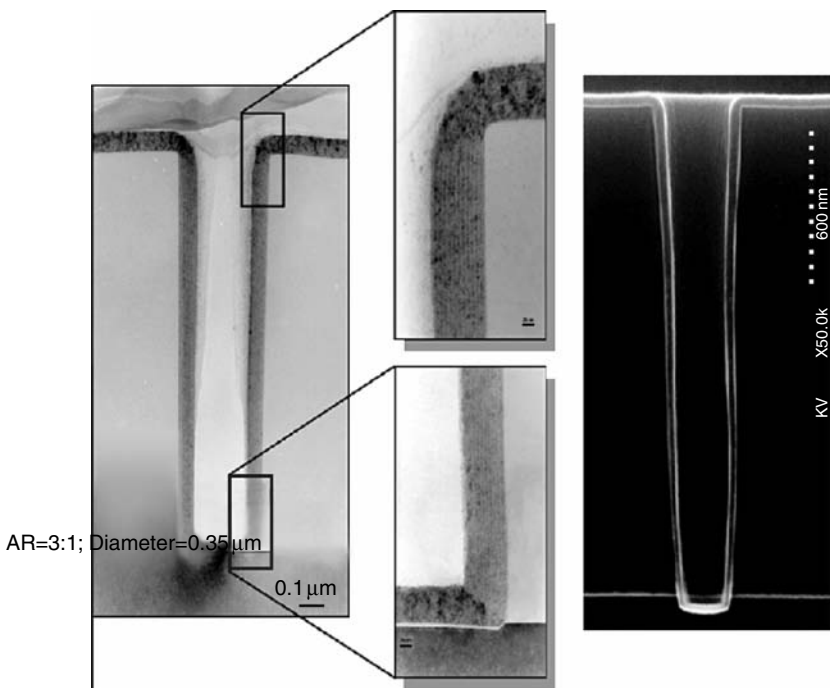


FIGURE 13.63 TEM for plasma treated MOTiN.

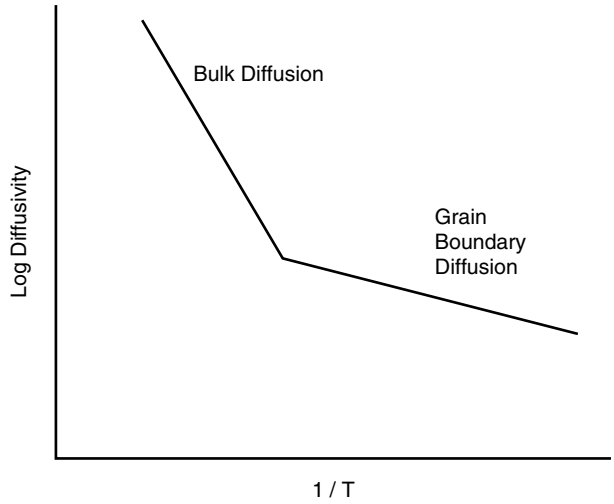


FIGURE 13.64 Diffusivity vs. 1/*T*.

decomposition layer thickness and plasma treatment power and time, the resulting film structure can vary from very fine grained, which is close to amorphous, to full film thickness-sized grains.

For barrier applications, density, and structure are the two most important characteristics other than material compatibility. Typical diffusivity as a function of temperature is shown in Figure 13.64.

At high temperatures, bulk diffusion through the grains dominates; at lower temperatures, grain boundary diffusion or short circuit diffusion becomes dominant. The ideal barrier with very low diffusivity should be dense and amorphous so that diffusion species cannot easily diffuse through the bulk film, along grain boundaries or defects.

13.4.2.5 ALD Transition and Refractory Metal Nitride

Transition metal nitrides are extensively used in microelectronic devices as diffusion barriers to prevent inter-diffusion and reaction between metals and silicon or dielectric materials. For example, titanium nitride (TiN) has been used as the barrier material to prevent fluorine diffusion to reaction with titanium, aluminum, and silicon. In addition, ALD TiN is evaluated as low resistive electrode to replace poly-Si for mid-gap metal gate in high performance logic devices and for deep trench electrode in DRAM applications. Tantalum nitride (TaN) is used as barrier to prevent copper from diffusing into dielectric material. In addition, TaN/Ta stack film has been evaluated as one of most possible candidate for NMOS metal gate application in 65 nm and 45 logic devices. Atomic layer deposition TaN/Cu stack film has been successfully demonstrated as midgap metal gate material. Furthermore, both TiN and TaN are widely used in microelectronic devices as adhesion promoters between metal and dielectric materials.

Extensive researches have been carried out in the deposition of ALD transition metal nitride using metal chloride and NH_3. The deposition temperature is in the range between 300 and 500°C. Chlorine concentration decreases by increasing deposition temperatures. The film resistivity in the range between 200 and 3000 μΩ-cm has been reported for TiN, NbN, and MoN deposition, where low resistive films are obtained by increasing deposition temperature. The only exception is on TaN deposition where only N-rich Ta_3N_5 phase is obtained. Experimental results show that the reducing power of both NH_3 and dimethyl hydrazine are not strong enough to reduce Ta(V) in $TaCl_5$ into Ta(III). Highly conductive TaN ($\sim$1000 μΩ-cm) has been deposited using zinc as an additional reducing agent. However, zinc tends to dissolve into the film and out-diffuse in the subsequent thermal processing steps, which leads to major concerns on the application for Si-based microelectronic devices.

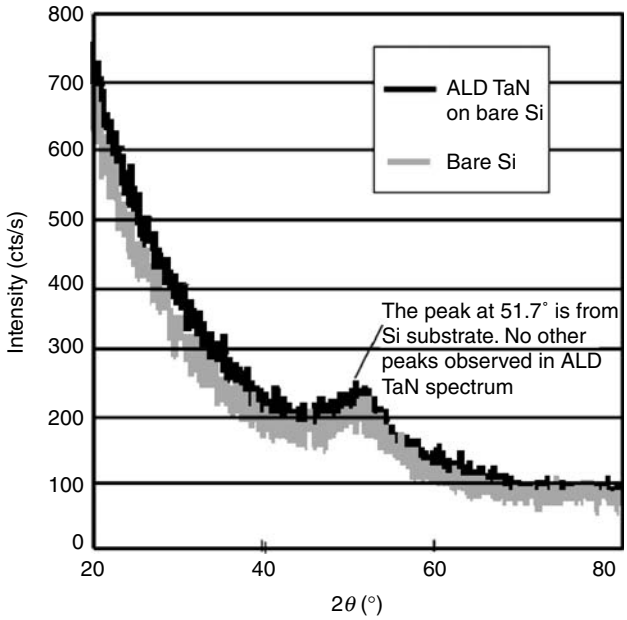


FIGURE 13.65 XRD spectrum recorded from 50 Å ALD TaN deposited using a MO precursor.

Many works have been done recently on the use of organo-metallic precursors for transition metal nitride deposition to prevent the risk of corrosion effects from halide compounds. The process temperature is typically below 300°C to prevent significant precursor thermal decomposition. Atomic layer deposition TaN deposited using Ta alkyl amides and NH_3 is a self-limiting process at less than 300°C. The film is of amorphous structure with excellent conformality. Figure 13.65 is a glancing incident XRD spectrum recorded from ALD TaN showing an amorphous structure. In addition, excellent Cu barrier property with film thickness below 20 Å has been reported. The major impurity is carbon with a concentration of well below than 5 at.%. Comparing with transition metal nitride deposited by halide compounds, the films deposited by organo-metallic precursors show higher film resistivity. For example, ALD TiN deposited using TDMAT/NH_3 chemistry at 200°C shows a film resistivity of ~5000 μΩ-cm, where a low deposition temperature is chosen to prevent TDMAT thermal decomposition. The high resistivity of as-deposited films is attributed to an amorphous film structure associated with a low deposition temperature. Post deposition plasma treatments significantly reduce the film resistivity and convert the films into nano-crystalline structures. Recent studies suggests that both *tert*-butylamine ($(CH_3)3CNH_2$, ^{t}BuNH$_2$) and allylamine (CH_2=$CHCH_2NH_2$, allylNH$_2$) are stronger reducer than NH_3 for TaN deposition using $TaCl_5$. Conductive TaN has been deposited by replacing NH_3 with both amines.

A new approach for metal nitride deposition is reported by plasma-enhanced atomic layer deposition (PEALD). Extensive works have been reported on depositing TaN using ^{t}BuNTa $(NEt_2)_3$ (*tert*-butylimidotris (diethylamido) tantalum) and atomic hydrogen. The deposition temperature is at 250°C. Comparing to thermal ALD TaN films, PEALD yields low resistive TaN (~400 μΩ-cm). The growth rate is approximately 0.08 nm/cycle. According to Auger electron spectroscopy analysis the films are Ta-rich and the carbon content in the films is about 15 at.%. The step coverage is comparable to thermal ALD processes. Figure 13.66 shows step coverage of ultra thin ALD TaN deposited by a thermal and (b) plasma enhanced ALD processes. Table 13.13 summarizes the film types and chemistries examined for transition metal nitride deposition.

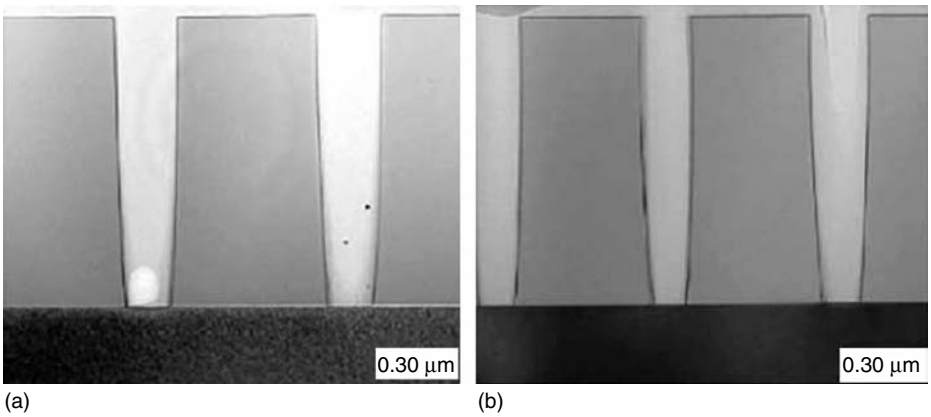


FIGURE 13.66 TEM images recorded from (a) thermal and (b) plasma enhanced ALD processes. Highly conformal step coverage is achievable by both processes.

13.4.2.6 Aluminum CVD

Chemical vapor deposition aluminum has just emerged for a promising semiconductor application. Compared with the standard PVD method, CVD Al from traditional precursor has suffered from higher resistivity, greater impurity content, and rougher surface, with a more difficult process and higher cost. Recent work has shown that CVD Al (using DMAH) in combination with PVD Al in an integrated process sequence under high vacuum, at temperatures less than 400°C, can fill $>8{:}1$ aspect ratio contacts without voids or seams (Figure 13.67). There is potential for electromigration advantages of void free filling at relatively low temperature in comparison with other aluminum filling methods or reactive ion etching patterning (Table 13.14).

Aluminum CVD was pioneered with tri isobutyl aluminum chemistry. Its deposition temperature is around 300°C–350°C, but the film has suffered from significant amounts of carbon [152–154]. The lower temperatures (200°C–250°C) of DMAH chemistry prevent carbon incorporation and provide excellent

TABLE 13.13 Film Types and Chemistries for Transition Metal Nitride Deposition

Film Type	Metal Precursor	Reducing Agent
TiN	$TiCl_4$, TiI_4	NH_3, $(CH_3)_2NNH_2$, NH_3+Zn
		$(CH_3)3CNH_2$, $CH_2{=}CHCH_2NH_2$
	$Ti(NMe_2)_4$, $Ti(NMeEt)_4$	NH_3
$TiSi_xN_y$	$Ti(NMe_2)_4$	NH_3+SiH_4
	$TiCl_4$	H_2+N_2 plasma $+SiH_4$
$TiAl_xN_y$	$TiCl_4$	$NH_3+Al(CH_3)_3$
Ta_3N_5	$TaCl_5$	NH_3, $(CH_3)_2NNH_2$
	$Ta(NMe_2)_5$, $Ta(NMeEt)_5$	NH_3
TaN	$TaCl_5$	NH_3+Zn, $(CH_3)3CNH_2$
		$CH_2{=}CHCH_2NH_2$
	^{t}BuNTa$(NEt_2)_3$	H_2 plasma
$TaSi_xN_y$	$TaCl_5$	H_2+N_2 plasma $+SiH_4$
$TaAl_xN_y$	$TaCl_5$	$NH_3+Al(CH_3)_3$
WN	WF_6, $(^t BuN)_2(Me_2N)_2W$	NH_3
NbN	$NbCl_5$	NH_3, $(CH_3)_2NNH_2$, NH_3+Zn
		$(CH_3)3CNH_2$, $CH_2{=}CHCH_2NH_2$
MoN	$NbCl_5$	NH_3, $(CH_3)_2NNH_2$, NH_3+Zn
		$(CH_3)3CNH_2$, $CH_2{=}CHCH_2NH_2$

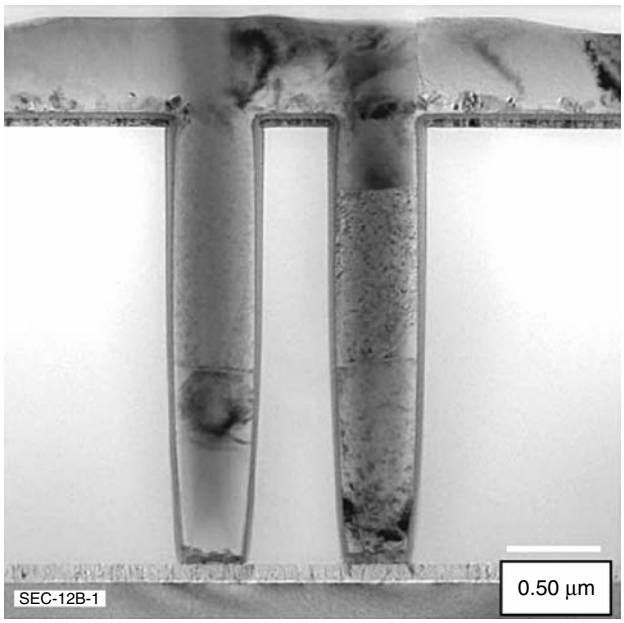


FIGURE 13.67 High aspect ratio contact fill by CVD Al/PVD Al integration.

stepcoverage. The purity and stepcoverage of DMAH CVD Al opens up a possible application as the wetting layer for aluminum flow: once a continuous Al layer is formed, under certain PVD conditions (350°C–450°C, slow deposition rate), surface energy will drive sputtered Al atoms to migrate into highly curvatured surfaces such as inside a contact hole, to smooth it and eventually fill it [155–157].

The DMAH chemistry has been postulated as a dimer or trimer form on the gas phase and on the surface where the methyl group and hydrogen are exchanged between precursors to yield TMA compound and CH_4 as by-products [158,159]. The monomethyl group bound to aluminum seems to be avoided at low temperature since its carbon content is extremely low (comparable with PVD Al).

TABLE 13.14 ALD Metal Films and Deposition Chemistries

Film Type	Metal Precursor	Reducing Agent
Al	$Al(CH_3)_3$	H_2 plasma
Ti	$TiCl_4$	H_2 plasma
Ta	$TaCl_5$	H_2 plasma
W	WF_6	Si_2H_6, SiH_4, B_2H_6
Mo	$MoCl_5$	Zn
Co	$Co(^i\text{Pr-Meamd})_2$	H_2
Ni	$Ni(acac)_2$, $Ni(^i\text{Pr-Meamd})_2$	H_2
Cu	CuCl	H_2, Zn, H_2 plasma
	$Cu(thd)_2$	H_2
	$Cu(acac)_2$	H_2
	$[Cu(^i\text{Pr-Meamd})]_2$	H_2
Ru	$Ru(Cp)_2$, $Ru(Od)_3$, $Ru(EtCp)_2$	O_2
	$Ru(EtCp)_2$	NH_3 plasma
Pt	$Pt(acac)_2$	H_2

acac: 2,4-pentanedione (acetylacetonate); od: 2,4-octanedionate; amd: acetamidinate; thd: 2,2,6,6-tetramethyl-3,5-heptanedione; Cp: cyclopentadienyl; ^{i}Pr: isopropyl.

At deposition temperatures greater than 300°C, carbon is found as an impurity and leads to higher resistivity [160].

Surface sensitivity is very predominant in the DMAH chemistry. Selective growth on clean Si and metal surfaces has been reported. The nucleation of aluminum on silicon dioxide is very sparse and slow, resulting in significant incubation time delay and very rough films. The control of selectivity is difficult and limits the acceptance of a selective process [161,162].

In a more realistic application, the silicon dioxide surface has to be desensitized by freshly deposited titanium, titanium nitride or metal-organic deposited titanium nitride, even it is loaded with hydrocarbon. Dimethyl aluminum hydride decomposition on the desensitized surface provides a thin continuous aluminum layer which is essential for the subsequent reflow process, especially for very small, high aspect ratio contact holes. The underlayer materials and crystal structure dictate the CVD aluminum film's preferred grain orientation and reflectivity [163,164]. For example, ⟨111⟩ orientation in the PVD Ti film will be duplicated in the following CVD Al [165,166]; a much more random orientation will result from metal-organic deposited titanium nitride.

13.4.2.7 Copper CVD

Copper metallization is a topic of great interest, for logic devices, and especially for microprocessors with clock speeds up to and beyond 1 GHz. RC delay using the conventional aluminum/silicon dioxide combination begins to exceed transistor delay as design rules drop to 0.18–0.15 μm. Current main stream process sequence is dielectric (SiO_2) deposition/trench, via etch/Cu diffusion barrier, Cu seed layer deposition/Cu electroplating. The Cu barrier/seed layer deposition is leading by PVD method. Chemical vapor deposition Cu provides a potentially more capable seed layer for electroplating Cu and further potential one step filling capability, especially at high aspect ratio via holes.

The precursors for CVD Cu are categorized as β-diketonated Cu compounds. The most widely reported is Cu(I)(hfac)(TMVS) [167–169]. The disproportionation reaction of Cu(I) to Cu and Cu(II) is the fundamental driving force:

$$2\text{Cu(I)(hfac)TMVS} \rightarrow \text{Cu(0)} + \text{Cu(II)(hfac)}_2 + 2\text{TMVS}$$

The reaction follows the standard model of supply limited regime to surface reaction controlled regime when precursor rate or wafer temperature are varied. The challenges of this chemistry are the stability of the liquid precursor, the volatility of the precursor and its by-products [170,171].

All the Cu precursors suffer from a similar dilemma: vapor pressure vs. decomposition temperature. In order to push higher deposition rate while maintaining good step coverage, the precursor supply rate should be as high as possible. To vaporize a large amount of liquid, heat supply rate and distribution becomes critical. With too much heat, the precursor self-decomposes before reaching the wafer. With too little heat, the precursor remains in the liquid state, causing instability of process, poor adhesion, poor uniformity, etc. The by-product $Cu(II)(hfac)_2$ is a low vapor pressure solid at room temperature. It can condense on cool surfaces: inside the deposition chamber, or in the foreline or exhaust line. To improve the precursor stability, an additional ligand component was blended into the precursor.

The deposition typically is executed at 180°C–220°C, 1 Torr pressure and 0.5–1 mg/m supply rate. The deposited Cu exhibits very good properties: <2 μΩ-cm at 5000 Å thickness, low tensile stress and low impurity content. The film's conformality is sufficient to fill >5:1 aspect ratio contact holes at the low end of the temperature spectrum. This enables electroplating to fill >10:1 trenches with only 500 Å as a seed layer. The integration of CVD Cu into devices is still in the development stage. Two potential candidates are: as a seed layer of high aspect ratio ($\geqq 5$:1) damascene via or contact for Cu electroplating (Figure 13.68); and as a wetting layer of warm PVD Cu reflow similar to the Al case.

13.4.2.8 ALD Metal

Atomic layer deposition metal processes are highly desired for many applications. For example, ALDW and Cu can be used as seed layers for CVD tungsten and ECP Cu fill processes. Ti and Ta are widely used as liner materials for Al and Cu interconnect metallization. In addition, metal films such as Ti, Ta, Ru, Pt,

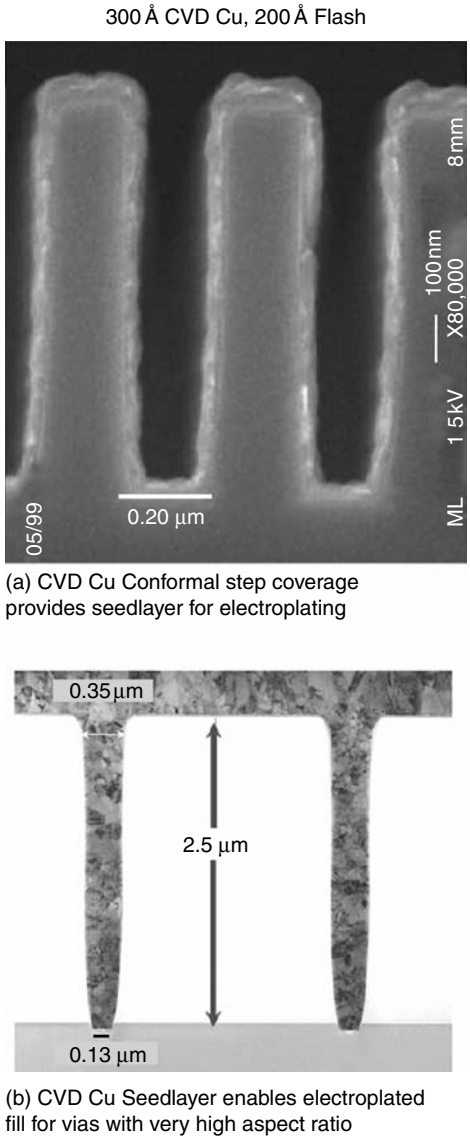


FIGURE 13.68 (a) Step coverage of CVD Cu seed layer and (b) filling by electroplating.

Mo, Ni, Co, W, and Cu are examined as candidates for low resistive electrodes for metal gate application. In addition, $CoSi_2$ and NiSi are replacing tungsten silicide for high speed, strained Si CMOS devices. Although ALD metals are highly demanded in many aspects, the process is an extremely challenging task from chemistry point of view with very limited success so far. The major difficulties encountered for ALD metal processes are

- Precursor stability against decomposition and disproportionation;
- Precursor volatility;
- Availability of effective reducing agent;
- Substrate dependency;
- Ultra-thin film continuity (nucleation).

13.4.2.8.1 Cu, Co, Ni and Pt

Atomic layer deposition Cu deposition was demonstrated by CuCl–H_2, Cu(thd)$_2$–H_2, and Cu(acac)$_2$–H_2 reactions. In all these cases, the growth process occurs only on appropriate metal surfaces. Since molecular hydrogen is very inert, the growth process occurs only if the substrate actively participates in the reaction. For example, in the growth of Cu by CuCl and H_2 chemistry, CuCl molecule is reduced by Ta substrates in the initial growth stage such that Cu is deposited and assists in the dissociation of molecular hydrogen into reactive hydrogen to facilitate the reaction. Similarly, the growth of ALD Cu, Cu(thd)$_2$ and H_2 reaction require a thin Pt/Pd alloy seed layer as a catalyst to dissociate Cu(thd)$_2$ to initiate the reaction.

Alternatively, ALD Cu deposition has been demonstrated from CuO reduction using Cu(thd)$_2$ and Cu(acac)$_2$. CuO is first deposited from the reaction of Cu(thd)$_2$ and Cu(acac)$_2$ with O_3 at a temperature below 150°C. The growth rates of CuO are 0.08–0.13 Å/cycle and 0.35–0.4 Å/cycle, respectively. This followed by CuO reduction using alcohols, aldehydes, carboxylic acids, H_2 or H_2-plasma. A temperature of 250°C–350°C is required by using alcohols or H_2 for CuO reduction. In contrast, H_2-plasma can be performed at approximately 150°C. The film resistivity is <10 μΩ-cm for 350 Å Cu film. The oxygen and carbon concentrations are below the detection limits by x-ray photoelectron spectroscopy. The use of atomic hydrogen generated by a remote plasma source reacts with Cu(acac)$_2$ has been reported recently. The process is performed at 140°C with H_2 plasma reduction time of 1–4 s The growth rate is approximately 0.3 Å/cycle with film resistivity of 10 μΩ-cm for 300 Å Cu film. The substrates applied for Cu deposition include silicon (–OH and –H terminated surfaces), ALD Al_2O_3, SiO_2, TiN, TaN, SiLK, and copper. Good adhesion is obtained on Cu deposited on all types of substrate investigated.

Similar to Cu deposition, Ni and Pt can be deposited using Ni(acac)$_2$ and Pt(acac)$_2$ reacting with H_2 on selected metal surfaces. In addition, metallic Ni deposition from NiO reduction is also reported. In this process, NiO is deposited using H_2O or O_3 as oxidants to react with Ni(acac)$_2$, followed by forming gas anneal. All the processes are performed at 250°C.

The deposition of metallic Co, Ni, and Cu using homoleptic *N*,*N*′-dialkylacetamidinato metal compounds and molecular hydrogen is reported recently. The growth processes are performed between 180 and 350°C. The growth rates varied from 0.5 Å/cycle for Cu to $\sim$0.1 Å/cycle for Co and $<$0.05 Å/cycle for Ni.

13.4.2.8.2 Emerging CVD/ALD Metals Applications

In future devices there are several potential areas that require new conducting materials deposited by CVD or ALD, implying new materials with good step coverage. One area is further refining copper diffusion barrier and seed layer for better compatibility with future porous low *k* dielectric materials; and alternative copper seed layer for enhancing copper grain size and maintaining low copper line resistance in small geometry. Another area is metal gate, an appropriate work function and process compatibility with potential high *k* dielectrics materials hafnium oxide, hafnium silicate, and Al_2O_3. A third area is capacitor electrode, compatible with BST and other ferroelectric or piezoelectric materials (SBT, PZT). This requires good oxygen compatibility, and should achieve low capacitor leakage and high capacitance.

References

1. Revesz, A. G. "The Defects Structure of Grown Silicon Dioxide Films." *IEEE Trans. Electron Device* ED-12 (1965): 97.
2. Wolf, S., and R. Tauber. *Silicon Processing for the VLSI Era*. Vol. 1. Sunset Beach, 198, CA: Lattice Press, 1986.
3. Kingery, W. D., H. K. Bowen, and D. R. Uhlmann. *Introduction to Ceramics*, 100. New York: John Wiley, 1976.
4. Cramer, J. K., and S. P. Muraka. *J. Appl. Phys.* 73, no. 5 (1993): 2458.

5. Yoshitomi, T., M. Saito, H. Oguma, Y. Akasaka, M. Ono, H. Nii, Y. Ushiku, H. Iwai, and H. Hara. *VLSI Symposium*, 99. 1993.
6. Xia, L.-Q., E. Yieh, S. Venkataraman, and B. C. Nguyen. *J. Electrochem. Soc.* 144 (1997): L117.
7. Takemura, H., S. Ohi, M. Sugiyama, T. Tashiro, and M. Nakamae. *Tech. Dig. Inst. Elect. Dev. Meet.* (1987): 375.
8. Robles, S., E. Yieh, and B. C. Nguyen. *J. Electrochem. Soc.* 142 (1995): 581.
9. Poenar, D., P. J. French, R. Mallée, P. M. Sarro, and R. F. Wolffenbuttel. *Sens. Actuators A* 41–42 (1994): 304.
10. Burt, D. L., R. F. Taraci, and J. E. Zavion. Motorola, Inc. U.S. Patent 3,934,060, 1973.
11. Wang, D. N., J. M. White, K. S. Law, C. Leung, S. P. Umotoy, K. S. Collins, J. A. Adamik, I. Perlov, and D. Maydan. Applied Materials, Inc. U.S. Patent 4,872,947, 1988.
12. Santoro, C. J., and D. L. Tolliver. *Proc. IEEE* 59 (1971): 1403.
13. Kern, W., J. L. Vossen, and G. L. Schnable. *11th Annu. Proc. Reliab. Phys. (Symp.)* (1973): 214.
14. Vegel, R. H., S. R. Butler, and F. J. Feigel. *J. Electron. Mater.* 14 (1985): 329.
15. Levin, R. M., and K. Evans-Lutterodt. *J. Vac. Sci. Technol.* B1 (1983): 54.
16. Becker, F. S., D. Pawlik, H. Anzinger, and A. Spitzer. *J. Vac. Sci. Technol. B* 5 (1987): 1556.
17. Ikeda, Y. NEC Corporation, U.S. Patent 5,462,899, 1995.
18. Kim, E. J., and W. N. Gill. *J. Electrochem. Soc.* 141 (1994): 3463.
19. Xia, L.-Q., E. Yieh, P. Gee, F. Campana, and B. C. Nguyen. *J. Electrochem. Soc.* 144 (1997): 3209.
20. Xia, L.-Q., S. Nemani, M. Galiano, S. Pichai, S. Chandran, E. Yieh, D. Cote, R. Conti, D. Restaino, and D. Többen. *J. Electrochem. Soc.* 146, no. 3 (1999): 118.
21. Gorczyca, T. B., and B. Gorowitz. "PECVD of Dielectrics." In *VLSI Electronics Microstructure Science*, edited by N. Einspruch. Vol. 8, 69–76. New York: Academic Press, 1984.
22. Rosler, R. S. "Techniques of Low Pressure CVD." *Semicond. Int.* (1984): 72.
23. Bohr, S., U. Ahmed, L. Brigham, R. Chau, R. Gasser, R. Green, W. Hargrove., et al. *IEDM Tech. Dig.* (1994): 273–6.
24. Chatterjee, A., M. Mason, K. Joyn er, D. Rogers, D. Mercer, J. Kuehne, and A. Esquivel. *SPIE* 2875 (1996): 39.
25. El-Kareh, B. *Fundamentals of Semiconductor Processing Technologies*, 72. Boston, MA: Kluwer Academic Publishers, 1995.
26. Appels, J. A., E. Kooi, M. M. Paffen, J. J. H. Schatorjé, and W. H. C. G.Verkuylen. *Philips Res. Repts.* 25 (1970): 118–32, 119.
27. Kwok, K., E. Yieh, S. Robles, and B. C. Nguyen. *J. Electrochem. Soc.* 141 (1994): 2172.
28. Kishimoto, K., M. Susuki, T. Hirayama, Y. Ikeda, and Y. Numasawa. *J. Electrochem. Soc.* 141 (1992): 149.
29. Elbel, N., Z. Gabric, W. Langheinrich, and B. Neureither. *VLSI Symposium*, 1998.
30. Gorowitz, B., T. B. Gorczyca, and R. J. Saia. "Applications of PECVD in VLSI." *Solid State Technol.* (1985): 197.
31. Dobkin, D. M., S. Mokhtari, M. Schmidt, A. Pant, L. Robinson, and A. Sherman. *J. Electrochem. Soc.* 142 (1995): 2332.
32. Fujino, K., Y. Nishimoto, N. Tokumasu, and K. Maeda. *J. Electrochem. Soc.* 140 (1993): 2922.
33. Robles, S., K. Russell, M. Galiano, V. Siva, V. Kithcart, and B. C. Nguyen. *J. Electrochem. Soc.* 143 (1996): 1414.
34. Thakur, R. P. S., F. Gonzalez, R. Hawthorne, V. Ward, and N. Jeng. *Mater. Res. Soc. Symp. Proc.* 133 (1993): 283.
35. Hsieh, J. C., Y. K. Fang, C. W. Chen, N. S. Tsai, M. S. Lin, and F. C. Tseng. *IEEE Trans. Elect. Dev.* 41 (1994): 458.
36. Xia, L.-Q., R. Conti, M. Galiano, F. Campana, S. Chandran, D. Cote, D. Restaino, and E. Yieh. *J. Electrochem. Soc.* 145, no. 5 (1999): 1884.
37. Imai, S., Y. Yabuuchi, Y. Terai, T. Yasui, C. Kudo, I. Nakao, and M. Fukumoto. *Appl. Phys. Lett.* 60, no. 22 (1992): 2761.
38. Xia, L.-Q., R. Conti, M. Galiano, F. Campana, S. Chandran, D. Cote, D. Restaino, and E. Yieh. *Electrochemical Society 193rd Meeting*, San Diego, CA, 1998.

39. Cramer, J. K., S. P. Murarka, K. V. Srikrishnan, and W. Patrick. *J. Appl. Phys.* 73, no. 5 (1993): 2458.
40. McCaughan, D. V., R. A. Kushner, and S. Wagner. *J. Electrochem. Soc.* 121 (1974): 724.
41. Pan, Y., L. Chan, and R. Sundaresan. Chartered Semiconductor Manufacturing Pte Ltd. U.S. Patent 5,742,088, 1997.
42. Saito, M., T. Yoshitomi, H. Hara, M. Ono, Y. Akasaka, H. Nii, S. Matsuda., et al. *IEEE Trans. Elect. Dev.* 40, no. 12 (1993): 3264.
43. Mayumi, S., and S. Ueda. *J. Appl. Phys.* 29, no. 4 (1990): 645.
44. Watanabe, H., S. Ohnishi, I. Honma, H. Kitajima, and H. Ono. *J. Electrochem. Soc.* 142, no. 1 (1995): 237.
45. Levin, R. M. *J. Electrochem. Soc.* 129 (1982): 1766.
46. Srinivasan, A., S. Sharan, and G. Sandhu. Micron Technology, Inc. U.S. Patent 5,731,235, 1998.
47. Yokozawa, A. NEC Corporation, U.S. Patent 5,663,087, 1997.
48. Muller, K. P., B. Flietner, C. L. Hwang, R. L. Kleinhenz, T. Nakao, R. Ranade, Y. Tsunashima, and T. Mii. *IEDM Tech. Dig.* (1996): 507.
49. Wang, Y., M. Li, A. Khan, K. Li, and S. Pan. *AVS National Symposium*, San Jose, 1997.
50. Doo, V. Y., D. R. Nichols, and G. A. Silvey. *J. Electrochem. Soc.* 113 (1966): 1279.
51. Arizumi, T., T. Nishinaga, and H. Ogawa. *Jpn. J. Appl. Phys.* 7 (1968): 1021.
52. Watanabe, T., N. Goto, N. Yasuhisa, T. Yanase, T. Tanake, and S. Shinozaki. *IEEE, IRPS 1987 Tech. Dig.* 50, (1987).
53. Ohji, Y., T. Kusaka, I. Yoshida, A. Hiraiwa, K. Yagi, and K. Mukai. *IEEE IRPS 1987 Tech. Dig.* 55, (1987).
54. Gabriel, C. VLSI Technology, Inc. U.S. Patent 5,294,295, 1994.
55. Wu, S. Powerchip Semiconductor Corporation, U.S. Patent 5,679,601, 1997.
56. Bohn, P. W., and R. C. Manz. *J. Electrochem. Soc.* 132 (1981): 85.
57. Nguyen, V. S. Electrochemical Society Extended Abstracts 83-1. 216. 1983.
58. Wristers, D. J., H. J. Fulford, and D. L. Kwong. Advanced Micro Devices, Inc. U.S. Patent 5,674,788, 1997.
59. Benedict, J. P., D. M. Dobuzinsky, P. L. Flaitz, E. N. Hammerl, H. Ho, J. F. Moseman, H. Pal, S. Yoshida, and H. Takato. IBM Corporation, U.S. Patent 5,763,315, 1998.
60. Peters, L. *Semicond. Int.* September (1998): 64.
61. Singer, P. *Semicond. Int.* November (1994): 52.
62. Ting, C., and T. E. Seidel. In *MRS Symposium Proceedings*. Vol. 381, 3, 1995.
63. Ip, F., and C. Ting. In *MRS Symposium Proceedings*. Vol. 381, 135, 1995.
64. Hendricks, N. H. *Solid State Technol.* July (1995): 117.
65. M'saad, H., M. Vellaikal, L. Zhang, Y. Wang, D. Witty, K. Rossman, and F. Moghadam. In *Proceedings of 5th International Dielectric for ULSI Multilevel Interconnection Conference*, 210. 1999.
66. Yang, S., et al. In *Proceedings of IEDM*, 197. 1998.
67. Homma, T., R. Yamaguchi, and Y. Murao. *J. Electrochem. Soc.* 140 (1993): 687.
68. Lee, P. W., S. Mizuno, A. Verma, H. Tran, and B. Nguyen. *J. Electrochem. Soc.* 143 (1996): 2015.
69. Albrecht, M. G., and C. Blanchette. *J. Electrochem. Soc.* 145 (1998): 4019.
70. Waeterloos, J., H. Meynen, B. Coenegrachts, T. Gao, J. Grillaert, and L. Van den Hove. In *Proceedings of 3rd International Dielectric for ULSI Multilevel Interconnection Conference*, 310. 1997.
71. Chua, C. T., G. Sartar, and X. Hu. *J. Electrochem. Soc.* 145 (1998): 4007.
72. Kohl, A. T., R. Mimma, R. Shick, L. Rhodes, Z. L. Wang, and P. A. Kohl. *Electrochem. Solid-State Lett.* 2 (1999): 77.
73. McClatchie, S., K. Beekmann, and A. Kiermasz. In *Proceedings of 4th International Dielectric for ULSI Multilevel Interconnection Conference*, 311, 1998.
74. Sahli, S., M. A. Djouadi, S. Hadj-Moussa, F. Mansour, and M. S. Aida. *Mater. Chem. Phys.* 33 (1993): 106.
75. Solid State Technology, (in preparation).
76. Grill, A., V. Patel, S. A. Cohen, D. C. Edelstein, J. R. Paraszczak, and C. Jahnes. In *Proceedings of 1996 MRS Meeting*, 417. Pittsburgh, PA, 1997.
77. Grill, A., V. Patel, and C. Jahnes. *J. Electrochem. Soc.* 145 (1998): 1649.

78. Matsubara, Y., K. Kishimoto, K. Endo, M. Iguchi, T. Tatsumi, H. Gomi, T. Horiuchi, et al. In *Proceedings of IEDM*, 1998.
79. Limb, S. J., D. J. Edell, E. F. Gleason, and K. K. Gleason. *J. Appl. Polym. Sci.* 67 (1998): 1489.
80. Labelle, C. B., S. J. Limb, and K. K. Gleason. In *Proceedings of 3rd International Dielectric for ULSI Multilevel Interconnection Conference*, 98, 1997.
81. Lu, T. M., J. A. Moore, J. F. MacDonald, C.-I. Lang, and G.-R. Yang. *CVD Technology, SEMI*, 39, 1995.
82. Gaynor, J. F., J. J. Senkevich, and S. B. Desu. *J. Mater. Res.* 11 (1996): 1842.
83. Treichel, H., et al. "Deposition, Annealing and Characterization of Tantalum Pentaoxide Films." *Mat. Res. Soc. Symp. Proc.* 282 (1993): 557–68.
84. Kwon, K. W., et al. "Thermally Robust Ta_2O_5 Capacitor for the 256-Mbit DRAM." *IEEE Trans. Electr. Dev.* 43, no. 6 (1996): 919–23.
85. Colombo, L. *Presented at ISIF 1998*, Monterey, CA, 1998.
86. Kingon, A. *Presented at ISIF 1999*, Colorado Springs, CO, U.S.A., 1999.
87. Eguchi, K., et al. In *Submitted to ECS for Proceedings of 193rd Meeting of ECS*.
88. Kamins, T. I. "Structure and Properties of LPCVD Si Films." *J. Electrochem. Soc.* 126 (1979): 833.
89. Meyerson, B. S., and W. Olbricht. "Phosphorous-Doped Poly-Si via LPCVD." *J. Electrochem. Soc.* 131 (1984): 2361.
90. Kamins, T. I. *Polycrystalline Silicon for Integrated Circuit Applications*. Norwell, MA: Kluwer Academic Publications, 1988.
91. Green, M. L., and R. A. Levy. *J. Met.* 37, no. 6 (1985): 63.
92. Ohba, T., M. Shirasaki, N. Misawa, T. Suzuki, T. Hara, and Furumura. In *Proceedings of VMIC, IEEE*, 226, 1990.
93. Sharan, S., and G. S. Sandu. *Advanced Metallization and Interconnect Systems for ULSI Application in 1997*, 23. Pittsburgh, PA: MRS Publishers, 1997.
94. McConica, C. M., and K. Krishnamani. *J. Electrochem. Soc.* 133 (1986): 2524.
95. Van Der Putte, P. *Tungsten and Other Refractory Metals for VLSI Application II*, 77. Pittsburgh, PA: MRS Publishers, 1986.
96. Bryant, A. *J. Electrochem. Soc.* 125 (1978): 1534.
97. Broadbent, E. K., and C. L. Ramiller. *J. Electrochem. Soc.* 131 (1984): 1427.
98. Blumenthal, R., G. C. Smith, H. Y. Liu, and H. L. Tsai. *Tungsten and Other Refractory Metal for VLSI Applications IV*, 65. Pittsburgh, PA: MRS Publishers, 1988.
99. Clark, T., A. P. Constant, M. Chang, and C. Leung. *Tungsten and Other Advanced Metals for ULSI/VLSI Applications V*, 167. Pittsburgh, PA: MRS Publishers, 1989.
100. Rana, W. S., J. A. Taylor, L. H. Holschwandner, and N. S. Tsai. *Tungsten and Other Refractory Metal for VLSI Applications II*, 187. Pittsburgh, PA: MRS Publishers, 1986.
101. Hara, T., T. Ohba, H. Yagi, and H. Tsutikawa. *Advanced Metallization for ULSI Application in 1993*, MRS.
102. Ellwanger, R., S. Ghanayem, and A. Mak. Applied Materials Internal Work.
103. Ohba, T. *Advanced Metallization for ULSI Applications*, 25. Pittsburgh, PA: MRS Publishers. 1991.
104. Green, M. L., Y. S. Ali, T. Boone, B. A. Davidson, L. C. Feldman, and S. Nakahara. *Tungsten and Other Refractory Metal for VLSI Applications II*, 85. Pittsburgh, PA: MRS Publishers, 1986.
105. Ng, S. L., S. J. Rosner, S. S. Laderman, T. I. Kaminsm, D. R. Braobury, and J. Amano. *Tungsten and Other Refractory Metal for VLSI Applications II*, 93. Pittsburgh, PA: MRS Publishers, 1986.
106. Rutten, M., D. Greenwell, S. Luce, and R. Dreves. *Advanced Metallization for ULSI Applications*, 277. Pittsburgh, PA: MRS Publishers, 1991.
107. Kobayashi, N., S. Iwata, N. Yamamoto, and N. Hara. *Tungsten and Other Refractory Metal for VLSI Applications II*, 159. Pittsburgh, PA: MRS Publishers, 1986.
108. Murarka, S. P. *Silicide for VLSI Applications*. New York: Academic Press, 1983.
109. Brors, D. L., J. A. Fair, K. A. Monnig, and K. C. Saraswat. *Solid State Technol.* April (1983): 183.
110. Brors, D. L., J. A. Fair, K. Monnig, and K. C. Saraswat. *Semicond. Int.* May, (1984).
111. Saraswat, K. C., D. L. Brors, J. A. Fair, K. A. Monnig, and R. Beyers. *IEEE Trans. Electr. Dev.* ED-30, no. 11 (1983): 1497.

112. Ellwanger, R. C., K. D. Prall, D. R. Malinaric, R. W. Williams, J. E. J. Schmitz, and E. I. Bromley. *Tungsten and Other Advanced Metals for VLSI/ULSI Applications Materials Research Society*, 335. Pittsburgh, PA: Materials Research Society, 1990.
113. Fukumoto, M., and T. Ohzone. *Appl. Phys. Lett.* 50, no. 14 (1987): 894.
114. Shioya, Y., S. Kawamura, I. Kobayashi, M. Maeda, and K. Yanagida. *J. Appl. Phys.* 61, no. 11 (1987): 5103.
115. Wright, P. J., and K. C. Saraswat. *IEEE Trans. Electr. Dev.* 36, no. 5 (1989): 879.
116. Wu, T., R. S. Rosler, B. C. Lamartine, R. B. Gregory, and H. G. Tompkins. *J. Vac. Sci. Technol.* B6, no. 6 (1988): 1707.
117. Srinivas, D., G. Raupp, and J. Hillman. *Tungsten and Other Advanced Metals For VLSI/ULSI Applications V*, 407. Pittsburgh, PA: Materials Research Society, 1990.
118. Inamdar, A. S., and C. M. McConica. *Tungsten and Other Advanced Metals For VLSI/ULSI Applications V*, 93. Pittsburgh PA: Materials Research Society, 1990.
119. Hillman, J., J. B. Price, B. Triggs, and M. Aruga. *Tungsten and Other Advanced Metals for VLSI/ULSI Applications 1990*, 329. Pittsburgh, PA: Materials Research Society, 1991.
120. Telford, S. G., M. Eizenberg, M. Chang, A. K. Sinha, and T. R. Gow. *J. Electrochem. Soc.* 140, no. 12 (1993): 3689.
121. Gow, T. R., R. J. Lebel, J. Schmitz, R. Chow, E. Bromley, L. Reed, and P. Arnold. *Advanced Metallization for ULSI Applications*, 557. Pittsburgh, PA: Materials Research Society, 1991.
122. Choi, D. K., and C. G. Ko. *J. Korean Institute Telematics Electron.* 29A, no. 9 (1992): 15.
123. Sharan, S., and G. S. Sandu. In *Proceedings of Workshop on Advanced Metallization and Interconnect Systems for ULSI Applications in 1997*, 23. Pittsburgh, PA: MRS Publishers, 1998.
124. Oshita, Y., and K. Watanabe. *J. Electrochem. Soc.* 145, no. 7, (1998).
125. Arena, C., J. Faguet, R. F. Foster, J. T. Hillman, F. Martin, and Y. Morand. In *Proceedings of Workshop on Advanced Metallization for ULSI applications in 1994*, 259. MRS Publishers: Pittsburgh, PA, 1994.
126. Hillman, J. T., R. F. Foster, J. Faguet, R. Arora, M. S. Ameen, C. Arena, and F. Martin. *Proc. VMIC, IEEE* (1994): 365.
127. Kubat, P., and P. Engst. *Appl. Surf. Sci.* 64 (1993): 97.
128. Alexandrescu, R., R. Cireasa, B. Dragnea, I. Morjan, I. Voicu, and A. Andrei. *J. Physique IV* 3 (1993): 265 Supplement to J. Physique II.
129. Xing, G. C., and M. C. Ozturk. *Mater. Lett.* 17 (1993): 379.
130. Akahori, T., T. Murakami, and Y. Morioka. In *Proceedings of VMIC, IEEE*, 405. 1993.
131. Oshita, Y., K. Watanabe, K. Tsuda, and T. Takada. In *Proceedings of Workshop Advanced Metallization and Interconnect Systems for ULSI Applications in 1997*, 685. Pittsburgh, PA: MRS Publishers.
132. Wolf, S. Silicon Processing for the VLSI Era. Vol. 2. Sunset Beach, CA: Lattice Press, 1990.
133. Wittmer, M. *Appl. Phys. Lett.* 37 (1980): 540.
134. Arora, R. D. Shrinivas, R. F. Foster, J. T. Hillman, and D. W. Studiner. In *Proceeding of Workshop on Advanced Metallization for ULSI Applications VIII*, 281. Pittsburgh, PA: Materials Research Society, 1993.
135. Lee, M.-B., H.-D. Lee, B.-L. Park, U.-I. Chung, Y. B. Koh, and M.-Y. Lee. *IEDM Tech. Dig.* (1996): 683.
136. Buiting, M. J., A. F. Otterloo, and A. H. Montree. *J. Electrochem. Soc.* 138 (1991): 500; Buiting, M. J., A. F. Otterloo, and A. H. Montree. *J. Electrochem. Soc.* 139 (1992): 2581.
137. Yokoyama, N., K. Hinode, and Y. Homma. *J. Electrochem. Soc.* 138 (1991): 190.
138. Srinivas, D., J. T. Hillman, W. M. Triggs, and E. C. Eichman. *Advanced Metallization for ULSI Applications*, 319. Pittsburgh, PA: MRS Publishers, 1991.
139. Hegde, R. I., R. W. Fiordalice, E. O. Travis, and P. J. Tobin. *J. Vac. Sci. Technol.* 11 (1993): 1287.
140. Hilton, M. R., L. R. Narasimham, S. Nakamura, M. Salmeron, and G. A. Somorjai. *Thin Solid Film* 139 (1986): 247.
141. Sandu, G. S., S. Meikle, and T. T. Doan. *Appl. Phys. Lett.* 62 (1983): 240.
142. Littau, K. A., R. Mosely, M. Eizenberg, H. Tran, A. Sinha, G. Dixit, M. K. Jain, M. F. Chisholm, and R. H. Havemann. In *Proceedings of VMIC, IEEE*, 440. 1994.
143. Fix, R. M., R. G. Gordon, and D. M. Hoffman. *Mater. Res. Soc. Symp. Proc.* 168 (1990): 357.

144. Raaijmakers, I. J. In *Proceedings of IEEE VMIC*, 260. 1992.
145. Foules, G. W. A., and F. H. Pollard. *J. Chem. Soc.* (1953): 2588.
146. 20C TDMAT vapor pressure is 0.05 Torr and TDEAT is 0.001 Torr, Schumacher data sheet.
147. Eizenberg, M., K. Littau, A. Mak, Y. Maeda, M. Chang, and A. K. Sinha. *Appl. Phys. Lett.* 65 (1994): 2416.
148. Eizenberg, M., K. Littau, S. Ghanayem, M. Liao, R. Mosely, and A. K. Sinha. *J. Vac. Sci. Technol.* A13 (1995): 590.
149. Konecti, A. J., G. A. Dixit, J. D. Luttmer, R. H. Havemann, M. Danek, and M. Liao. *Proc. IEEE VMIC* (1996): 181.
150. Iacoponi, J., M. Liao, M. Danek, K. Littau, D. Saigal, M. Eizenberg, and R. Mosely. In *Proceedings of MRS Conference ULSI XI*, 375. 1996.
151. Liao, M., M. Danek, and D. Smith. Applied Materials Internal Report.
152. Bent, B. E., R. G. Nuzzo, and L. H. Dubois. *J. Am. Chem. Soc.* 111 (1989): 1634.
153. Bent, B. E. *J. Vac. Sci. Technol.* A6, no. 3 (1920): 1988.
154. Egger, K. W. *J. Am. Chem. Soc.* (1969): 2869.
155. Guo, T., L. Y. Chen, D. Brown, P. Besser, S. Voss, and R. Mosely. *Thin Solid Film* 332 (1998): 319.
156. Clevenger, L. R., et al. *Proc. IITC, IEEE* May (1998): 137.
157. Guo, T., et al. *MRS Spring Meeting Symposium K*, March 1997.
158. Littau, K. A., R. Mosely, S. Zhou, H. Zhang, and T. Guo. *Microelectron. Eng.* 33 (1997): 101.
159. Willis, B. G., and K. F. Jensen. *Advanced Metallization and Interconnect Systems for ULSI Application*, 29. 1996.
160. Kawamoto, H., H. Sakaue, S. Takehito, and Y. Horiike. "Japanese." *J. Appl. Phys.* 29, no. 11 (1990): 2657.
161. Sugai, K., H. Okahayashi, T. Shinzawa, S. Kishida, A. Kobayashi, T. Yako, and H. Kadokawa. *J. Vac. Sci. Technol.* B13, no. 5 (1995): 2115.
162. Kondoh, E., and T. Ohta. *J. Vac. Sci. Technol.* A13, no. 6 (1995): 2663.
163. Avinun, M., et al. *J. Appl. Phys.* (in press).
164. Avinun, M., M. Naik, T. Guo, R. Mosely, K. Littau, S. Zhou, and L. Chen. *Thin Solid Film* 320 (1998): 67.
165. Naik, M., L. Chen, T. Guo, R. Mosely, I. Beinglass, and F. Chen. *Proc. VMIC Conf.* 97 (1997): 383.
166. Chen, L. Y., M. Naik, T. Guo, R. Mosely, F. Chen, and I. Beinglass. *MRS Spring Meeting Symposium K*, March 1997.
167. Norman, J. A. T., and B. A. Muratore. U.S. Patent No. 5,085,731, 1992.
168. Girolami, G. S., P. M. Jeffries, and D. L. Dubois. *J. Am. Chem. Soc.* 115 (1993): 1015.
169. Jain, A., T. T. Kodas, and M. J. Hampden-Smith. *J. Electrochem. Soc.* 140 (1993): 1434.
170. Peterson, G. A., J. E. Parmeter, C. A. Apblett, M. F. Gonzales, P. M. Smith, T. R. Omstead, and J. A. T. Norman. *J. Electrochem. Soc.* 142 (1995): 939.
171. Gelatos, A. V., R. Marsh, M. Kottke, and C. J. Mogab. *Appl. Phys. Lett.* 63 (1993): 2842.

14

Atomic Layer Deposition

14.1 Introduction **14**-1

14.2 ALD Origins **14**-3

14.3 Chemical Processes **14**-5
Basic Sequential Self-Limiting Processes • Illustration of ALD Chemical Reactions by Cartoons • Types of ALD Reactions • Temperature Dependence of Thermal ALD • Saturation Characteristics • Throughput Calculator • Digital Control, Linearity with Cycling • Initiation Processes • Nanolaminates • Rapid ALD by Limited Reactions

14.4 ALD System Technology **14**-20
General Description • ALD Systems

14.5 Applications **14**-25
Higher-*K* Oxide Capacitors on Chip • Advanced Dielectrics and Metal Gates

14.6 Summary of Current Status and Outlook **14**-32

Acknowledgments **14**-33

References **14**-33

Thomas E. Seidel
AIXTRON, Inc.

14.1 Introduction

Atomic layer deposition (ALD) is a variant of chemical vapor deposition (CVD) technology. Conventional CVD uses a continuous supply of reactants that co-exist in space and time above the wafer substrate. Chemical vapor deposition may have chemical reactions in the gas phase or on the surface. *Atomic layer deposition is carried out using sequential exposures of chemical reactants, each reactant having self-limiting depositions separated in time and space. In ALD, chemical reactions take place only on the surface*. The self-limiting surface reacting feature assures conformal film coatings even with very high aspect ratio structures. Each pair of sequential chemical reactions has a thickness granularity at the atomic level, typically depositing approximately 0.1 to several Å/pair of chemical exposures. Atomic layer deposition is by definition digital, adding discrete increments of thickness as the film growth proceeds.

Atomic layer deposition technology has existed for about 35 years [1], however, it was not widely utilized, mainly due to throughput (TP) limitations. Atomic layer deposition made its first appearance as a potential solution in the semiconductor industry's International Technology Roadmap for Semiconductors (ITRS) Roadmap in the last several years [2]. One reason for the current emergence of ALD within the semiconductor applications community is that the ALD capability now intersects the needs for current and future device scaling. In particular, the need of conformal coatings on high aspect ratio structures with thickness approximately 10's of Å were defined for capacitor dielectrics with aspect ratios

TABLE 14.1 Capacitor

Year of Production	2005	2010
DRAM Product	1G	4G
Structure	SIS	MIM
Aspect Ratio	19	94
Dielectric	$SiON/Al_2O_3$	high-K

ranging from approximately 20–100:1 at feature sizes (FSs) below 100 nm. It was only when the needs of the semiconductor industry and ALD capability were aligned that a widespread interest developed in ALD technology for semiconductor applications.

In this Introduction section, we review selected elements of the ITRS Roadmaps for DRAM capacitors, gates, and interconnects. Those needs are related to the potential solutions offered by ALD. More detailed discussions of Roadmap elements are developed in Section 14.5.

General scaling elements for *Capacitors* are shown in Table 14.1. The extreme aspect ratios range from 19 to 94. They are for the cylinder/stack in 2005 and the trench in 2010, respectively. Currently, ALD Al_2O_3 is in production using cylinder/stack architecture. Beyond 2007, other higher dielectric constant materials ("high-K's") are needed, such as HfO_2, ZrO_2, and nanolaminates or alloys with Al_2O_3. Both stack and trench capacitor structures are migrating from semiconductor–insulator–semiconductor (SIS) to metal–insulator–metal (MIM) structures. Conformal dielectrics and conformal metal electrodes, especially for the lower electrode, are needed. Atomic layer deposition is well suited for these applications. The high aspect ratio and the large active areas for capacitors are challenging requirements, but are within developing ALD capabilities.

Selected scaling elements for low power and high performance *Gates* are shown in Table 14.2. Complimentary metal-oxide silicon (CMOS) gates in production today use SiON dielectrics and are planar and migrating to fully depleted silicon on insulator (SOI). The leakage for SiON is rising below an equivalent oxide thickness (EOT) of approximately 20 Å, and much high-K work has been carried out to achieve both lower EOT values and lower leakage. Recently, ALD technology appears to have achieved parity with CVD and PVD techniques for the gate performance using polysilicon [3]. These and similar solutions are particularly promising for Lo power use, since certain EOT Roadmap requirements appear to be achieved in the laboratory. Yet, implementation of high ks with poly gates is limited by EOT depletion capacitance contributions and Fermi level pinning that shifts the threshold voltages to high values and is accompanied by channel mobility degradation [4]. As a result, solutions have moved to metal gates, which may alleviate the Fermi pinning and have shown promise of higher mobility on unstrained substrates. Metal gates provide for a reduced-EOT because of the reduction of the depletion width in the poly gate targeting EOTs below 1 nm by 2010. A necessary condition for achieving EOTs below 1 nm is the reduction of parasitic SiO_2 at the silicon surface.

The failure to develop a commercially released integrated high-mobility solution has led to the emergence of the use of strained silicon. It is now likely that higher Ks will be developed in combination with strained silicon, especially as strained silicon has come into widespread production. Beyond 2010, short channel effects have driven the gate architectures to 3D structures (e.g., Finfet-type architectures) and high Ks are likely to be used around the same time. Atomic layer deposition could be a good candidate because of conformal ALD film quality covering the edges of these gate structures.

TABLE 14.2 Gate

Year of Production	2005	2010
Structure	Planar	FD–SOI
EOT Lo power (nm)	1.5	0.9
EOT Hi perf (nm)	1.2	0.7

TABLE 14.3 Interconnect

Year of Production	2005	2010
Logic structure	DD, w ~8 levels	
Wiring AR intermed lev	1.6	1.7
DRAM contact AR	15	>20
Logic barrier thk. (nm)	10	5

Selected scaling elements for *Interconnect* are shown in Table 14.3. The interconnect requirements are challenged by the barrier thickness control in moderate aspect ratio features. The aspect ratios in the wiring interconnects that are used for logic are not scaling aggressively. This is because if aspect ratios were concurrently increased, then the capacitance between intermediate wiring levels would rise to unacceptably high levels. Atomic layer deposition is well suited to provide improvement in the scaling of the Cu wiring lines because of the controlled uniform and conformal thickness allowing the scaling of barriers below 7 nm.

Contact applications in stack/cylinder DRAMs, on the other hand, have high aspect ratio challenges that track the aspect ratios of the stack/cylinder DRAM capacitor. In 2010, the DRAM aspect ratio is greater than 20 and this is an opportunity for ALD TiN barrier and ALD W or other metal fill.

Aside from the three mainstream semiconductor applications (*Capacitor, Gate*, and *Interconnects*), other electronic segments such as thin film head sensors (data storage) are scaling to higher complexity using ALD. Atomic layer deposition technology has been commercially applied to "gap dielectrics" in thin film head sensor devices to provide high yield at densities of 40–80 GBit/in.^{2} [5]. The film thicknesses are typically approximately 10 nm. This enterprise provides additional critical mass for ALD technology since the requirements and equipment are, to a large extent, common to those of semiconductors.

There are also applications beyond classical CMOS for appliqué to the Nano-device era. However, before alternative advanced future memories are implemented; many innovations in scaling conventional DRAM using enhanced 3D approaches are being developed [6]. Once these initiatives are exhausted, advanced future memory devices may emerge. The ITRS mentions a number of such alternatives: magnetic RAM (MRAM), phase change devices, single-electron, and molecular memories. These memory architectures may initially have relaxed aspect ratio topologies relative to today's DRAM devices. However, they all require ultra-thin films with the inevitable architectural migration to moderate, if not extreme aspect ratios. Atomic layer deposition may be assessed on a case-by-case basis to determine the technology of choice for the advanced materials needed in these future technologies. For example, in the case of MRAM, ultra-thin approximately 10 Å tunnel dielectrics layers may use ALD because of its atomic level thickness precision.

This ALD chapter has content on: ALD Origins, Chemical Processes, ALD System Technology, Applications, and Summary of Current Status and Outlook.

14.2 ALD Origins

In the 1970s, Tuomo Suntola [1] and his co-workers at Lahja and then later at Microchemistry developed "atomic layer epitaxy (ALE)" and are widely recognized for many developments relating to early ALD. The commercialization of ZnS was demonstrated using sequential saturating chemical reactions of $ZnCl_2$ and H_2S. Atomic layer epitaxy included early commercial application to electroluminescent displays utilized, for example, at the Helsinki Airport, mid 1980's.

Today's ALD films have amorphous or polycrystalline structure, depending on the film composition and deposition conditions. The development of early commercial ALD tooling preceded the recent rapid growth for applications for semiconductors. Many ALD tutorials and classic articles have emerged out of

the Suntola [1] and the "Helsinki Group," which continues to make significant ALD technical and infrastructural contributions [7].

Also in the decade of the 1970s, a group led by Aleskovsky published "molecular layer deposition" methods [8]. A chemical ALD halide/hydride sequence with an exchange reaction was defined and nanolaminates of Al_2O_3 and TiO_2 were demonstrated. The vision for widespread application to future scaled micro-electronic devices was articulated.

There are many other alternative names for ALD, in addition to ALE and Molecular Layer Deposition, depending on the context of introduction [9].

Another early initiative using ALD was the pursuance of applications to 3–5 compounds, notably by Japanese researchers [10] and academic laboratories [11] in the1980s. In 3–5 compound researches, there has been little drive for high aspect ratio device architectures, so molecular beam epitaxy (MBE) with its directional nature but with its enabling vacuum instrumentation and impurity control became common for R&D. Metal organic chemical vapor deposition has emerged as the deposition technology of choice for commercial 3–5 applications.

In the last 5 years ALD progress has accelerated. Review articles by Ritala [9] and Kim [12] are recent key references. Equipment and processes have moved through two generations and now are moving to the third generation, to be hall-marked by higher productivity, reliability and other enhancements. The growth and importance of ALD technology is currently indicated by the commercial growth of ALD deposition equipment for use in semiconductor and data storage production. In the period of 2003–2004, approximately $100M equipment revenues were achieved [13]. See Figure 14.1. Revenues approaching $1B are projected by 2010. Future growth will be challenged by the attainment of improved precursor delivery and system productivity.

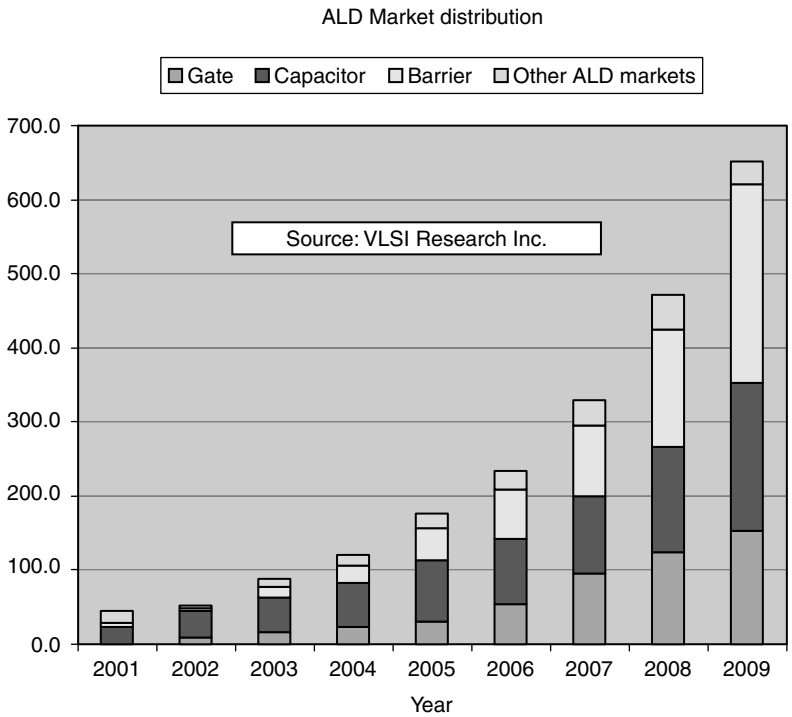


FIGURE 14.1 Past and forecasted atomic layer deposition (ALD) equipment markets. The *Y*-axis is the total ALD equipment revenues in units of $M. (From: Hutchenson, D., *VLSI Forecast*, VLSI Research, Inc., 2005.)

In summary, ALD technology provides a film deposition capability like no other. Fractions of atomic layers are deposited using discrete, self-limiting chemical reactions. The self-limiting depositions provide for continuous atomic layering on suitably prepared surfaces and provide inherently uniform coatings on high topology devices such as DRAM capacitors with design rules below 100 nm.

14.3 Chemical Processes

14.3.1 Basic Sequential Self-Limiting Processes

Atomic layer deposition processes take place in a low-pressure reactor at typical pressures in the range 0.1–5 Torr. The pressures are predetermined by the partial pressures of reactive precursors used in ALD just as in CVD processes. Higher pressure (e.g., approaching the atmospheric level) processes might be run with shorter exposure times, but their use would come at the expense of longer purge times. In ALD a first chemical precursor ("A") is pulsed, for example, bringing a metal species to the substrate surface in a first "half reaction." A first chemical precursor is selected so its metal reacts with suitable underlying species (e.g., O) to form new self-termination bonding and to provide a self-limiting deposition. Excess unused metal reactants and the reaction by-products are removed (e.g., as described by Bedair, [14] "…exposing the surfaces to …precursors in the absence of the (other) precursor…"). This is carried out by an evacuation-pump down and/or by entrainment with a flowing inert purge gas [8]. Then a second chemical precursor ("B") brings a non-metal such as active oxygen or nitrogen species to the surface, wherein the previously reacted passivating ligands of the first half reaction are reacted with new ligands from the second precursor, creating an "exchange by-product." A reaction of the non-metal takes place with the metal species to form a metal-oxide or metal nitride film. The second reactant also forms self-termination bonding with underlying reactive species to provide another self-limiting and saturating second half reaction. A second purge period is used. In ideal ALD, both the first and the second half reactions are self-limiting.

The chemistry of the ALD reactants must be suitable: the reactions must be fast, completely irreversible, and self-limiting. Removal of remnant precursors is to be achieved so that no vapor phase reactions occur, only ALD surface reactions. In the ideal, there is no physical adsorption of reactants or of the by-products. Finally, there should be no etching of the film by the reactants or the by-products.

The stoichiometry of films deposited by classic ALD (i.e., those in self-limited saturation) is driven by thermodynamics. It is actually rather difficult to achieve non-shoichiometric compounds deposited by ALD, although impurities may be incorporated. The value of the deposition per cycle is determined -in part- by the amount of reactants supplied, the number of molecules put into reaction, and not just the number of reactive sites.

A typical four stage ALD cycle is:

1. Expose A precursor for time (t_{ex1}) to carry out the first surface reaction
2. Removal time of the unused precursor and reaction products of reaction 1, t_{r1}
3. Expose B precursor for time (t_{ex2} to carry out the second surface reaction).
4. Removal time of the unused precursor and reaction products of reaction 2, t_{r2}.

These sequences are shown in Figure 14.2. The processes are repeated to build the film. The cycle time (CT) is defined as the sum of exposure and removal periods.

$$\mathrm{CT} = t_{ex1} + t_{r1} + t_{ex2} + t_{r2} \tag{14.1}$$

The CT can be as small as a fraction of a second or as long as a few minutes. The "removal" steps are often referred to as "purge" or "evacuation" steps. The removal step may be considered a non-value added step in the sense of not building the ALD film, but its efficiency and success limits parasitic CVD reactions. Furthermore, the removal of impurities such as halides or carbon may take place during the

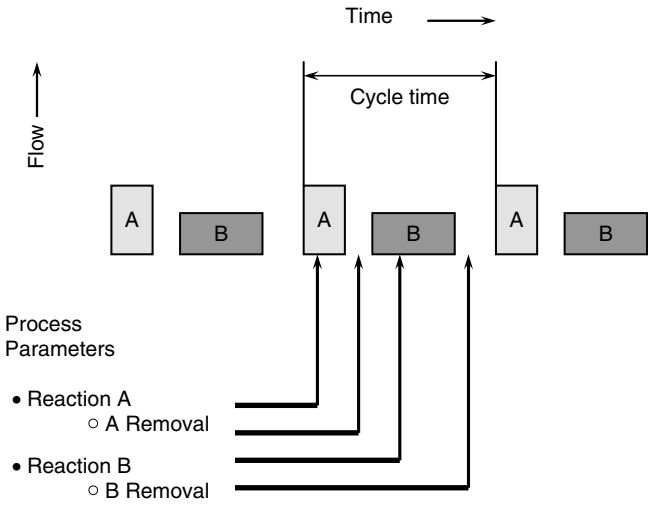


FIGURE 14.2 The four stage cycle: expose/removal/expose/removal of ALD. The cycle time is indicated.

removal step, so there is value in this regard. Any method for reducing the removal time without adverse effect, such as incurring significant parasitic CVD depositions, is desired.

14.3.2 Illustration of ALD Chemical Reactions by Cartoons

In this section, we illustrate an ALD reaction using the classical TMA/H_2O reaction chemistry [15]. In Figure 14.3a (left side), the surface is shown as initially "functionally activated" using hydroxylation (–OH). The –OH density may or may not be maximized. In Figure 14.3a (left side), the first half reaction proceeds as follows:

$$OH + Al(CH_3)_3 \rightarrow OAl(CH_3)_2 + CH_4 \tag{14.2}$$

After all of the active –OH sites have reacted with TMA, the reaction achieves saturation. The surface is now CH_3 terminated, which blocks or passivates the surface to any further reaction with TMA. See the right side of Figure 14.3a. The reaction actually takes place as a two-step chemical reaction process [16] with two separate activation energies; the TMA first adsorbs to the surface and interacts as a Lewis acid base complex with an exothermic energy of interaction. This is followed by an exchange reaction with additional exothermic component, where the H is removed as the by-product CH_4. The reaction is called "exchange" because the surface was terminated with –OH, but now is exchanged with $–CH_3$ terminations. In the ideal case, the by-product CH_4 molecules and trapped H (as part of a OH or HOH group) are not contained within the newly formed layer. In the right side of Figure 14.3b, we show trapped H in the surface region. The un-reacted TMA and byproducts CH_4 are shown leaving in the right side of Figure 14.3a and are removed during the first removal period, t_{r1}.

The passivasting $–CH_3$ ligands provide a new functionally activated surface for the next half reaction, assuming that the next exposure chemistry is energetically favorable. If water (HOH) is brought to the $–CH_3$ terminated surface, then $–CH_3$ is converted to CH_4 as a by-product and new –OH passivating ligands terminate the surface. The un-reacted H_2O and by-product CH_4 are shown leaving the right side of Figure 14.3b and are removed during second removal period, t_{r2}. This is illustrated in Figure 14.3b, with the second half reaction as follows:

$$OAl(CH_3)_2 + 2HOH \rightarrow OAl(OH)_2 + 2CH_4 \tag{14.3}$$

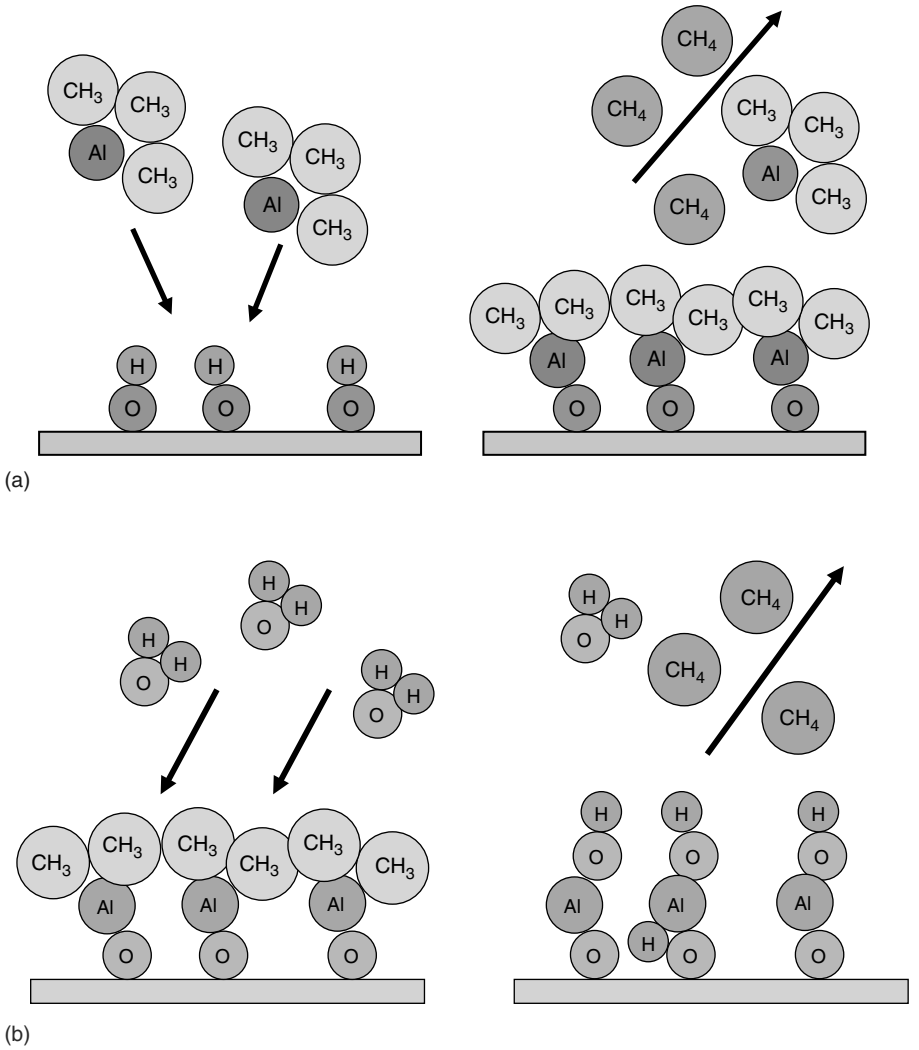


FIGURE 14.3 (a) Cartoon of reaction with OH (hydroxylated) surface exposed to TMA, followed by chemical reaction and the removal of CH_4 byproducts and unused TMA reactant. (b) Cartoon of reaction with a CH_3 terminated surface exposed to H_2O, followed by chemical reaction and the removal of CH_4 byproduct and unused H_2O reactant.

After all the HOH reacts with all accessible CH_3 sites, the reaction achieves saturation and the –OH act as passivating ligands to block the deposition of additional HOH reactant. Again, the reaction actually takes place as a two-step process [16], with two separate energies: the HOH first adsorbs to the surface and interacts as a Lewis acid base complex, and this is followed by an exchange reaction with additional exothermic component where the CH_3's are removed, again as the by-product CH_4. The surface was terminated with CH_3 but now is exchanged with OH terminations. The sequential process is repeated to build the desired film thickness.

The knowledge that each half reaction takes place in the manner described is supported by Fourier transform infrared (FTIR) surface science studies [17]. The FTIR signatures for $–CH_3$ decrease as the signatures from –OH increase for the H_2O half reaction and vice versa. In these studies, the FTIR of samples with boron nitride particles or porous silica surfaces are used to achieve high signal levels and to

determine the spectra of the terminating ligands, thus defining the time-dependent development and end state of the saturating half reactions.

In a typical ALD process the average thickness deposited per cycle is in the range of 0.1 to several Å/cycle, depending on the chemistries. It is often intuitively expected that ALD could give a layer thickness that is of the order the thickness of a lattice constant or about approximately 3 Å. However, this is rarely observed. The smaller thickness is attributed to the differences in the character and configurations of "steric effects." Passivating ligands take physical space and limit the ability of incoming reactants to reach and react with the underlying substrate below the passivating ligands. In the case of TMA/H_2O chemistry shown in Figure 14.3, the steric effects of $-CH_3$ passivating molecules are a rationale for less than a "full" Al_2O_3 monolayer per chemically exposed pair of reactants.

As an ideal case, we consider the number of atoms that could be in a monolayer ALD process. For Al_2O_3, the Atomic Density (Al_2O_3 molecules/cm^3) is given by:

$$\text{Atomic density} = (\text{Avogadro's No.})\left(\frac{\text{density}}{\text{g/mole}}\right)$$

$$= 6 \times 10^{23}(\text{molecules/mole})\frac{3(\text{g/cm}^3)}{102(\text{g/mole})} \sim 1.8 \times 10^{22}\text{molecules/cm}^3;$$

The number of Å/molecule is the cube root of the reciprocal of the atomic density approximately 3.7 Å. Aside, the surface density (#molecules/cm^2) is just (molecules/cm^3)$^{2/3}$ or approximately 7×10^{14} molecules/cm^2.

The two half reactions described above are:

$$OH + Al(CH_3)_3 \rightarrow AlO(CH_3)_2 + CH_4 \tag{14.4}$$

$$AlO(CH_3)_2 + 2HOH \rightarrow AlO(OH)_2 + 2CH_4, \tag{14.5}$$

and may be balanced and written for an overall TMA/H_2O reaction as:

$$2Al(CH_3)_3 + 3H_2O \rightarrow Al_2O_3 + 6CH_4 \tag{14.6}$$

14.3.3 Types of ALD Reactions

Atomic layer deposition may be carried out by thermal reactions or by plasma-assisted processes (which are also partially thermally activated). Various reactions are discussed. Although all the reactions described below are actually comprised of a two-step ALD process, single unbalanced heuristic chemical reactions are used unless the reaction chemistry is unclear without an explicit two-step description. The reader may refer to the original references to obtain greater clarity about detailed half reactions. Each of the reactions shown below is selected from important examples of films used in the development of ALD technology and applications.

14.3.3.1 Thermal ALD

14.3.3.1.1 Reactions Forming Compounds Using Halide–Hydride Chemistry

In this class of reactions, metal halides are reacted with hydrides, such as H_2O or NH_3. $TiCl_4$ and WF_6 are well known halide sources with good vapor pressures. However, many other metallic halides are often solids and may be sublimed at convenient source temperatures, but have low vapor pressures. A thermal ALD reaction of note using metal halides is the $HfCl_4$ chemistry:

$$HfCl_4 + 2H_2O \rightarrow HfO_2 + 4HCl \tag{14.7}$$

HfO_2 has recently emerged as a promising high-K dielectric because of its chemical stability with silicon relative to ZrO_2. A basic $HfCl_4/H_2O$ process has been demonstrated [18]. Other metal oxides with known halide reactions include: ZrO_2 [19], Ta_2O_5 [20], Al_2O_3 [21], and TiO_2 [22,23]. Additional references for these compounds may be obtained from the review of Ritala and Leskela [8].

TiN has been a standard material for barriers and contacts for over several decades [24]. A TiN process using $TiCl_4$ and NH_3 has been developed [25].

$$6TiCl_4 + 32NH_3 \rightarrow 6TiN + N_2 + 24NH_4CL \tag{14.8}$$

Other metal nitrides such as WN [26] and MoN reactions have been demonstrated with Cl or F-based precursors. TaN has been demonstrated as well although ALD's thermodynamic nature produces Ta_3N_5, which has a semiconducting level resisitivity. More reactive forms of N have been considered for reduction of metal halides, such as hydrazine [27]. However, due to the safety concerns of hydrazine, methylized derivatives, e.g., $(CH_3)_2NNH_2$, *t*BuNH2, have been reported for TiN, MoN, TaN, and NbN [28,29].

SiO_2 [30,31] and SiN [32] processes have been demonstrated using halide chemistry with H_2O and NH_3, respectively although with deposition rate limitations.

$$SiCl_4 + 2H_2O \rightarrow 2SiO_2 + 4HCl \tag{14.9}$$

$$Si_2H_2Cl_2 + NH_3 \rightarrow SiN + HCl \tag{14.10}$$

More recently metal organic alternatives are available for ALD SiO_2, as shown in Equation 14.16 below.

14.3.3.1.2 Reactions Forming Compounds Using Metal Organic (MO) Chemistry

While early ALD processes were often developed with the halide/hydride chemistry, the use of TMA for aluminum bearing films was an important exception. The lower vapor pressure of many metallic halide solids with concurrent particulate issues has led to the widespread use of liquid precursors for ALD. The excellent vapor pressure and thermal stability of TMA prompted the development of alternative MO liquid precursors for other metals. There are many existing MO's precursors, a few examples will be given in this emerging field.

Moving from halide sources to MO sources eliminates trace Cl or F, but introduces the presence of trace C or N. This is a necessary trade off. Organic chemistry provides a wide diversity of materials. Among other precursor characteristics, it is desired to develop liquid precursors with higher pressure *and concurrently* with higher thermal stability.

Thermal ALD reactions using metal organics include the following.

Al_2O_3, using the TMA/H_2O chemistry (Equation 14.6) [15] was the standard process for many years, but semiconductor device leakage characteristics were found to be better with TMA/O_3 chemistry [33].

$$Al(CH_3)_3 + HOH \rightarrow Al_2O_3 + CH_4 \tag{14.11}$$

and

$$Al(CH_3)_3 + O_3 \rightarrow Al_2O_3 + \text{H, C, O containing byproducts} \tag{14.12}$$

AlN (a III–V semi-insulating compound semiconductor) can be made with TMA/NH_3 [34], and may be considered for combination use with high-*K* insulators.

$$Al(CH_3)_3 + NH_3 \rightarrow AlN + 3CH_4 \tag{14.13}$$

High quality HfO_2 has been made using tetra-ethylmethylamino hafnium (TEMAH) [35–37]. Other metal alkyl amide MO compounds have been used in ALD reactions [38].

$$Hf[N(C_2H_5)(CH_3)]_4 + 2H_2O \rightarrow HfO_2 + 4NH(C_2H_5)(CH_3) \quad (14.14)$$

The use of O_3 as an oxidant has been demonstrated with TEMAH and other alkyl amides for HfO_2 [39–42]. Additionally, tris-dimethylaminosilicon (tDMAS) with O_3 has been demonstrated for the formation of SiO_2 for the use in the composite formation of $HfSiO_x$ [4,43,44].

$$Hf[N(C_2H_5)(CH_3)]_4 + O_3 \rightarrow HfO_2 + H, C, N, O \text{ byproducts} \quad (14.15)$$

$$HSi[N(CH_2)]_3 + O_3 \rightarrow SiO_2 + H, C, N, O \text{ byproducts} \quad (14.16)$$

Alkyl amides are also useful for the formation of nitrides. As an example, semi-insulating dielectric Hf_3N_4 is formed at nominally 200°C, and can transform to metallic HfN at ~1000°C [45].

$$Hf[N(C_2H_5)(CH_3)]_4 + NH_3 \rightarrow Hf_3N_4 + 12HN(C_2H_5)(CH_3) \quad (14.17)$$

14.3.3.1.3 Reactions Forming Elemental Films Using Halide–Hydride Chemistry

Alternating pulses of a silicon halide and silicon hydride can lead to the formation of elemental silicon [46].

$$SiCl_4 + Si_2H_6 \rightarrow 3Si + H_2 + 4HCl \quad (14.18)$$

14.3.3.1.4 Reactions Forming Elemental Films Using Metal Halides with Silane Reduction Chemistry

An early demonstration of the ALD formation of elemental W films has been reported [47].

$$WF_6 + SiH_4 \rightarrow W + SiF_xH_y + SiF_4 + H_2 \quad (14.19)$$

The strong bonding energy of the SiF_x by-product compound essentially leaves no Si to react with the metal, so WSi_x is not formed. This chemistry is similar to that known for CVD W processes. Other refractive metals can be formed using silane or other hydride-based reduction reactions.

14.3.3.1.5 Reactions Forming Noble Metals with O_2 Chemistry

This chemistry [48] may run counter to initial intuition, as oxygen "combustion" is used to create an elemental material. However, RuO is an intermediate. Once RuO_x forms on the substrate surface, it is further reduced by $Ru(Cp)_2$ to form an elemental material:

$$RuO_x + Ru(Cp)_2 \rightarrow Ru - Ru(Cp) + CO_2 + H_2O \quad (14.20)$$

$$Ru(Cp) + O_2 \rightarrow RuO_x + CO_2 + H_2O \quad (14.21)$$

where $(Cp)_2 = (C_5H_5)_2$. After each the O_2 half-reactions the surface is terminated in RuO_x. However, after each pair of half-reactions, an additional layer of Ru is added to the bulk film. The concentration of the O may have to be controlled to avoid oxidation of the underlying layers.

14.3.3.2 Plasma Assisted ALD

14.3.3.2.1 Reactions Forming Elemental Films Using Metal or Silicon Halides with Atomic H

An early Si elemental ALD thermal process used SiH_2Cl_2/H_2 [49], but the reaction proceeds above 800°C, which is too high for many applications. ALD saturation and elemental Si (or Ge) were attained using

atomic H for reduction at approximately 540°C. [50]. Still later, other researchers used the same reaction principle to deposit non-group IV elements, and made elemental Ti and Ta [12,51].

$$SiCl_4 + {}^*H \rightarrow Si-H + HCl;\ \text{repeat halide/}{}^*H\ \text{sequences} \tag{14.22}$$

$$TiCl_4 + {}^*H \rightarrow Ti-H + HCl;\quad \text{repeat halide/}{}^*H\ \text{sequences,} \tag{14.23}$$

where * indicates radical or plasma environment. Plasma assisted ALD may be by direct plasma, remote apparatus configurations or combinations thereof. Plasma excitations typically create a distribution of species, ions, radicals, and neutrals. These species in turn have distributions in energies and various electronic excitation states. A radical, strictly speaking (a free radical) is an atom or a molecule with an unpaired electron in the electronic level scheme, and therefore potentially very reactive as it may give or receive a single electron to complete a paired spin state. Examples are $^*H$ or $^*CH_3$ and ozone, which do not have an unpaired electron in its electronic molecular coordination, and therefore, strictly speaking is not a radical, but has often been called a radical because of its high reactivity.

Ta [12] may be deposited in this manner as well.

$$\text{Si surface} - OH + TaCl_5 \rightarrow TaCl_x + HCl \tag{14.24}$$

$$TaCl_x + {}^*H \rightarrow Ta-H + HCl \tag{14.25}$$

The first reaction is the initiation step on the Si surface. The second step builds the Ta film, after this the $TaCl_5/^*H$ sequence is repeated to build the film thickness.

It should be mentioned that plasma ALD reactions are typically carried out with only the non-metal in the plasma environment. If the metal precursor is placed in the plasma environment, it will decompose and lead to non-saturating deposition, just as in CVD. The preferred plasma assisted ALD mode, then, is to pulse the metal precursor in a thermal mode followed by its purge, followed in turn by plasma assisted non-metal reaction. Typically these reactions may occur approximately 100°C lower than their thermal counterparts.

14.3.3.2.2 Reactions Forming Metal Compounds Using Halide or MO Chemistries

Plasma assisted metal precursors use metal halides or MO compounds to make metal nitrides.

$$TiCl_4 + {}^*NH_3 \rightarrow TiN-H + HCl \tag{14.26}$$

The reaction takes place at approximately 100°C lower than its thermal counterpart. Metal nitrides or oxides may be formed using halide precursors and plasma containing oxidants (O_3, $^*H_2O\ldots$) or nitridants ($^*NH_3$) [12,52], as well as otherwise non-reactive gases O_2 and N_2/H_2.

Metal organic precursors and plasma-activated precursors can also be used. An example [53] is TBTDET:

$$(CH_3)_3CN \;=\; Ta[N(C_2H_5)_2]_3 + {}^*H \rightarrow TaN + \text{C, H, N byproducts} \tag{14.27}$$

14.3.3.3 Catalytic ALD (Enhancement of Reaction Rates and Thickness per Cycle for SiO_2)

A useful chemistry for ALD SiO_2 has been difficult to develop notwithstanding the recent *t*DMAS + O_3 process (see Equation 14.16 above). The classical halide/hydride SiO_2 process is slow. Two catalytic methods have emerged that address this issue.

14.3.3.3.1 Catalytic ALD for Reaction Rate Enhancement

$SiCl_4/H_2O$ reaction chemistry produces approximately 1.1 Å SiO_2/cycle around 400°C. The overall reaction may be limited by the sluggish kinetics of the OH half reaction with $SiCl_4$ [54]. The use of pyridine (C_5H_5N) as a continuous background catalyst during the $SiCl_4/H_2O$, ALD sequence promotes reaction down to less than 80°C and reduces the reactant dose by approximately five orders of magnitude.

Above 80°C, the pyridine desorbs and the reaction is not catalytically enhanced. The catalytically assisted ALD deposition rate of 1.35 Å/cycle is similar to the non-catalytic rate. The FTIR OH stretching frequency is shifted to the red during pyridine exposure, consistent with an energy reduction of the limiting reaction(s). See also Klaus [55].

$$SiCl_4 + HOH \xrightarrow{\text{w pyridine}} SiO_2 + HCl \tag{14.28}$$

14.3.3.3.2 "Catalytic Layer ALD" with Large Granularity Alternating Layer Chemistry

This process is initiated by a (few) catalytic Al_2O_3 layers using TMA/H_2O chemistry and is followed by a tris(tert-butoxy) silanol exposure. Silica film growth for this catalytic cycle saturates at about 100 Å/cycle, depending on the reaction temperature. A maximum deposition rate (Å/cycle) occurs at around 200°C. The process results in a nanolaminate, where a layer of a few Å of Al_2O_3 is under about 100 Å of SiO_2. The process can be repeated to build thicker films. The theory of this "giant" saturating chemistry is proposed to be related to a cross-linking network mechanism and is obviously fundamentally different than conventional ALD. The films have reasonable electrical properties. The developers of this deposition technology have used the general term "alternating layer deposition" referring to the alternation of TMA/H_2O and the SiO_2 [56]; it has also been called "Catalytic Layer ALD" because of the catalytic role of the TMA/H_2O.

While a fundamentally different saturating layer method of film deposition than classical ALD, it is not deposition at the atomic level because each cycle gives 20–100 Å. This process does not have two sequential chemical exchange reactions taking place. However, the chemistry does exhibit saturation, since the silanol step is self-limited after multiple layers have become cross-linked. Hence, it may be suitable for a high topology and high deposition rate class of applications.

$$Al(CH_3)_3 + HOH \rightarrow Al_2O_3 - H + CH_4 \tag{14.29}$$

$$Al_2O_3 - H\ (cat) + \text{silanol} \rightarrow Al_2O_3(\text{few}\%)/SiO_2 \tag{14.30}$$

14.3.4 Temperature Dependence of Thermal ALD

The idea of an "ALD Window," where the deposition rate (Å/cycle) is constant with varying deposition temperature has been discussed [1,11]. Atomic layer deposition processes are thermally activated, requiring sufficient thermal energy to make surface reactions occur, but not so large a thermal energy that the passivating ligand is unstable. Atomic layer deposition reactions that have a broad maximum may appear to be constant. This is shown in Figure 14.4. Atomic layer deposition mechanisms responsible for *lower* ALD deposition rates are:

1. At lower temperatures, there is insufficient energy to achieve a complete chemical reaction, and
2. At higher temperatures there is too much thermal energy for the net thermal stability of some of the molecules participating in one or both of the passivating ligands, for TMA/H_2O, e.g., the –OH has a desorption rate greater than its adsorption rate.

This general behavior is illustrated in the solid curve in Figure 14.4. Atomic layer deposition is generally a low temperature process, most processes having observed saturating reactions from 125 to 500°C. However, for a given chemistry, there may only be a 50°C–100°C nearly constant ALD deposition rate (Å/cycle) or "window."

In other words, ALD deposition processes are *thermally activated and temperature-dependent*. In practice, there is seldom a region of truly constant deposition rate over a wide temperature range. At low temperatures, the chemical adsorption-reaction dominates and the deposition rate increases with temperature. At higher temperatures, a region of desorption dominates and the deposition rate decreases with temperature.

The features of Figure 14.4 are as follows: a non-ALD deposition associated with condensation phenomena is shown at lower temperatures in upper left (dashed line) of the diagram. Additionally,

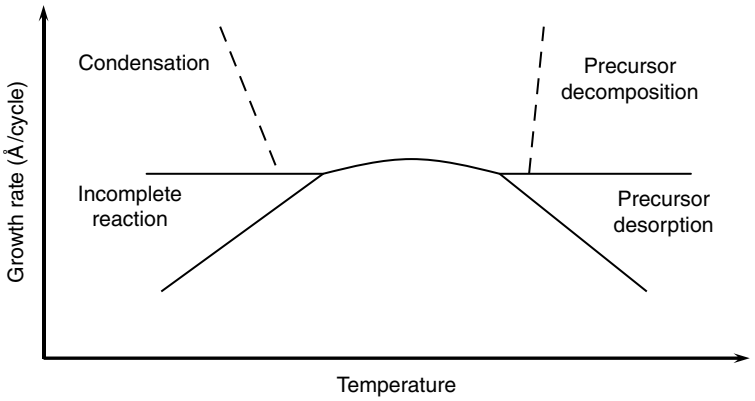


FIGURE 14.4 An illustration of the temperature dependence of the ALD deposition rate and related processes. At lower temperatures, incomplete ALD reactions and condensation effects are possible, and at high temperatures, ALD reactant surface desorption and precursor decomposition may take place.

deposition by pyrolytic CVD from the decomposition of precursors at higher temperatures is shown in the upper right (dashed line). These processes that may result in observed deposition rates *higher* than the ALD process. The onset of condensation or precursor decomposition effects may or may not occur at the corners of the ALD nominally constant deposition region.

The temperature dependence behavior is evident for the TMA/H_2O reaction. This is shown in Figure 14.5. Data for the temperature dependence of TMA/O_3 is also shown, where the maximum is not well defined, but may occur at lower temperatures.

The temperature dependence of the ALD deposition rates (Å/cycle) for Al_2O_3 using TMA and H_2O (triangles) and O_3 (dots) oxidants are compared [40]. Open circles are H_2O published data [15]. The single O_3 point (asterisk) result at 350°C is shown [33], where the electrical properties of O_3 were reported for the first time to be marginally, but importantly superior to that of H_2O-based Al_2O_3 films. The leakage was lower for O_3-based vs. H_2O-based chemistry. These and other results have led to the O_3 chemistry becoming the process of choice for DRAM capacitors. Reaction chemistry for TMA/O_3 has been proposed [41].

The preferred use for semiconductors is in the higher temperature region. The data storage thin film head application has a preferred temperature below 200°C. Different applications and different film use suggests the need to determine the electrical properties for different oxidants and temperatures for various films, such as HfO_2. For example, the ALD deposition rate (Å/cycle) for HfO_2 using tetra-ethylmethylamide hafnium (TEMAH), tetra-diethylmethyamide hafnium (TDMAH), and hafnium chloride $HfCl_4$ are shown in Figure 14.6 [37,40]. The physical characteristics such as crystal structure and refractive index also depend on the deposition temperature.

The HfO_2 deposition rates (Å/cycle) for H_2O and O_3 are shown. There is little difference in the ALD results for all the precursors used; however, both MO precursors decompose above 320°C. Consistent with the thermal stability of TEMAH being greater than TDMAH, the parasitic CVD of TEMAH is less than TDMAH. This is an illustration of the precursor decomposition and the parasitic CVD deposition effect shown in the upper right hand region of Figure 14.4. TEMAH has a higher thermal decomposition temperature than TDMAH, however, it has a lower vapor pressure.

14.3.5 Saturation Characteristics

The methodology and context of saturation phenomena is discussed. Saturation characteristics are considered as a proof that the chemistry under study is, in fact, being run in an ideal or near ideal ALD

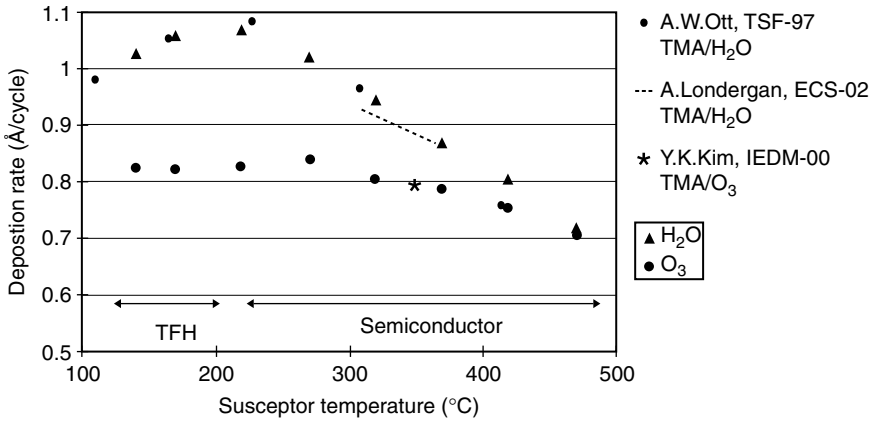


FIGURE 14.5 The ALD deposition rates (Å/cycle) for Al_2O_3, using TMA and H_2O (triangles) and O_3 (large dots). (After Seidel, T., et al., *American Vacuum Society ALD-03*, Abstr. Book/CD, San Jose, CA, 2003). These data are compared to the published H_2O data by Ott. (From Ott, A. W., *Thin Solid Films*, 292, (1997): 135.) (small dots), (From Londergan, A. R., et al., *Rapid Thermal and Other Short Time Processes*, Vol. III, ECS, 2002.) (dashed line), and (From Kim, Y. K., et al., *IEDM Tech. Dig.*, (2002): 369.) (asterisk).

mode. Yet, the behaviors of different chemistries are quite diverse. The most obvious consideration is that the reaction under consideration may not be fundamentally self-saturating. There are many variations on this, $TiCl_4/NH_3$, e.g., provides a good saturation for the $TiCl_4$ half reaction and poor saturation for the NH_3 [57]. In practice, it is often found that one of the half reactions is gradually or softly saturating, while the other is rapidly or more ideally saturating. If one of the reactions is self-limiting, it is empirically found that good step coverage may be expected, but it may not be as good as if both reactions show ideal saturation. However, in the studies of TiN [58], the TDMAT half reaction continues to show thickness increase with exposure time, while the NH_3 exhibits saturation. A variety of soft saturations for metal oxides using metal halides and H_2O chemistries have been studied [59].

The methodology for measuring the saturation characteristics is as follows: in a given reactor and under given pressure and temperature conditions, the precursor removal (purge) time and the "second"

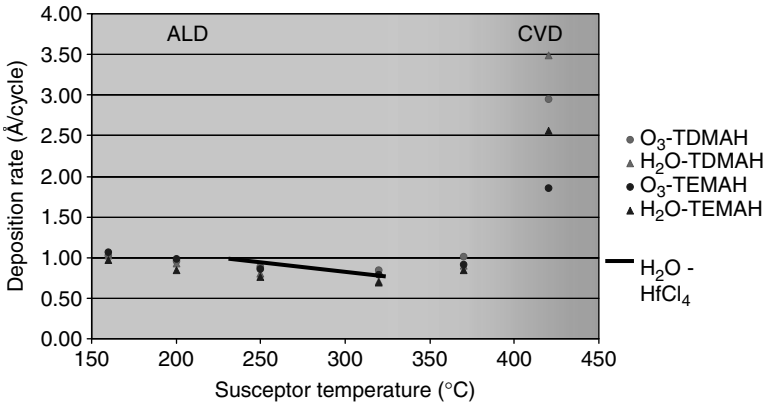


FIGURE 14.6 Data of the ALD deposition rate (Å/cycle) for HfO_2 using TEMAH and TDMAH with H_2O and O_3. The onset of chemical vapor deposition (CVD) from pyrolytic deposition from the decomposing precursors occurs at higher temperatures. (After Seidel, T., et al., *American Vacuum Society ALD-03*, Abstr. Book/CD, San Jose, CA, 2003.)

precursor exposure time are set arbitrarily large, e.g., 10 s. The "first" exposure time is systematically varied from a small time (e.g., 0.05 s up to several seconds), while running a standard number of cycles, e.g., 100 cycles. The resulting average film thicknesses/cycle for that number of cycles are plotted against the first exposure times (see Figure 14.7). The average thickness per cycle (Å/cycle) vs. the exposure time exhibits saturation. The process is repeated for the second reactant, holding the first exposure fixed at or above its predetermined saturation time. Then the purges can be reduced until significant overlap of the pulses exists and the onset of CVD is observed. One can accept purge times with a certain percentage of CVD that is permitted by the application, uniformity, and step coverage results. The minimum exposure times and removal times are added to determine the best operational CT (seconds).

If a variety of different detailed chemical reaction pathways are operative and each with different kinetics and reaction time constants, then saturation characteristic is the sum of these different processes although each reaction pathway may exhibit sharp saturations itself.

Alternately, the reactor may provide parasitic surfaces (always the case), which are more sticky to one of the precursors (H_2O or NH_3) and the possibility exists that these precursors are never sufficiently removed during the removal cycle. There is always a background CVD deposition component added to the pure ALD, and the longer the exposure the more additional total deposition takes place.

Another factor to be considered is the "reactivity" of the precursor. A more reactive precursor may react soon when arriving at the surface, while a less reactive precursor may undergo diffusion on the surface or may desorb before it reacts. Low reaction coefficient precursors may have long time constants.

The *observed* time constant for the saturation characteristic can be influenced by the capability of the chemical precursor delivery. In the case of a high vapor pressure source, such as TMA with 10 Torr at room temperature, the adequate delivery of precursor results in 200 or 300 mm wafers reaching saturation in approximately 0.1 s. However, lower vapor pressure sources can result in chemical source limited saturation time constants.

Still another factor includes byproducts that may etch the deposited film resulting in a slowing of apparent saturation that can be chemical concentration and dosage-dependent. As the reaction proceeds toward saturation, the thickness is *reduced* as time proceeds. This idea can conceptually lead to limiting thicknesses smaller than would be without the effect of self-etching of the by-products.

It may be difficult to separate these different mechanisms. Examples of gradual or softly saturating reactions are the H_2O half reaction with TMA and NH_3 with $TiCl_4$. In Figure 14.7, we show the ideal saturation of TMA (with H_2O). At a 100 ms exposure time, the ALD deposition rate is already saturated. The H_2O (with TMA) saturates of over the range up to 1000 ms [52,59].

The number of precursor molecules passed through a reactor and over the wafer can be calculated using the partial pressure of the precursor dose and the exposure time. Work by Ott [15] shows that 10^4–10^5 Langmuirs or 10–100 mTorr-s of precursor dose is needed to reach saturation in the Al_2O_3 process using TMA/H_2O (1 Langmuir $= 1 \times 10^{-6}$ Torr-s).

14.3.6 Throughput Calculator

The TP depends on the thickness of the film to be deposited, the ALD deposition rate (Å/cycle) and CT(s). It is useful to develop a "calculator" for TP wafers per hour (WPH).

$$\begin{aligned} &\text{TP(wph)} = 60 \text{ min/h/[\#min/wafer], or} \\ &\text{TP} = 60/[t_f(\text{Å})/\text{FDR(Å/min)} + t_{oh}(\text{min/waf})], \text{ or} \\ &\text{TP} = 60/[(t_f(\text{Å})/60 \text{ min/s} \times \text{ADR(Å/cycle)/CT(s)}) + t_{oh}], \end{aligned} \tag{14.31}$$

where t_f is the film thickness (Å) for a wafer and Film Deposition Rate (FDR) (Å/min) and ADR is the ALD deposition rate (Å/cycle). The ratio gives the raw value of the film deposition time (min/wafer). The wafer overhead time, t_{oh} is the handling and temperature equilibration time.

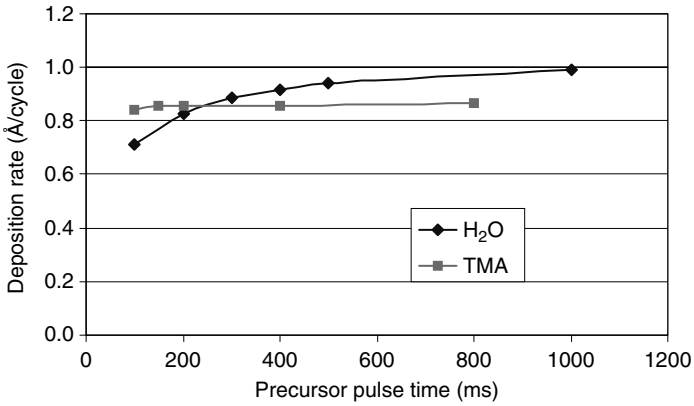


FIGURE 14.7 The saturation characteristics of TMA/H_2O. Note that H_2O is softly saturating. (After Londergan, A. R., et al., *Rapid Thermal and Other Short Time Processes*, Vol. III, 2002, ECS.)

For example, for an ALD deposition rate of 1 Å/cycle, and a CT of 3 s, the FDR is 60 s/min × 1 Å/cyc × cyc/3 s = 20 Å/min. Then for a film thickness t_f of 40 Å and a wafer overhead time, t_{oh}(handling and equilibration) of 2 min/wafer, the TP (WPH) is 60/[(40/20) + 2] = 15 wph. Thus, the single wafer TP has reasonable productivity for thin films (40 Å) of moderately good CTs (3 s).

The "FDR" (Å/unit time or Å/min) is an important intermediate parameter, which we will use again in an independent optimization. See Section 3.10, below on limited reactions.

14.3.7 Digital Control, Linearity with Cycling

The methodology for determining the deposition rate Å/cycle is described in this section. Figure 14.8 shows the digital increase in thickness as a function of the number of exposure cycle per reactions for TEMAH using H_2O (Figure 14.8a) and O_3 (Figure 14.8b), respectively [36]. The thickness increases linearly with the number of cycles. The slope corresponds to 0.81 Å/cycle. The measurements were taken on a calibrated ellipsometer. The result is extracted from the incremental additive thickness above the initial native oxide thickness of approximately 10 Å. The thicknesses were measured after approximately 60–150 cycles of exposure.

If one starts with a native oxide and makes the first thickness measurement, there will be an offset at zero cycles corresponding to the initial native oxide thickness. If the surface is prepared with an HF last process, with –H termination, there may be a delay in continuous layering until a few layers are deposited after which the curve is linear, see Section 3.8, below.

Such thickness vs. the number of ALD cycle plots are used to determine the reproducibility of ALD depositions and especially the linearity for the first few cycles, which are indicative of the success of ALD initiation processes. There is no reason to expect that the deposition rate on a substrate different from the film being grown (grown on an "unlike surface") will be the same as the deposition rate on a substrate of the same film being grown (grown on a "like surfaces.") We expect like-on-like depositions to give equal incremental additive thicknesses. However, nanolaminate interfaces where one switches from one material to another often have interface incubation effects.

It is noted that linearity of thickness vs. the number of ALD cycles is a characteristic of ALD process, but this characteristic cannot be used to distinguish ALD from "pulsed CVD." Pulsed CVD means both precursors are pulsed at the same time. If both the precursors are pulsed simultaneously, but the pulses are separated by a time period, then the additive thickness is linear with the number of cycles, but the process is pulsed CVD and not ALD. Conformality is dictated by device topology and the amount of surface reaction content relative to gas phase reaction.

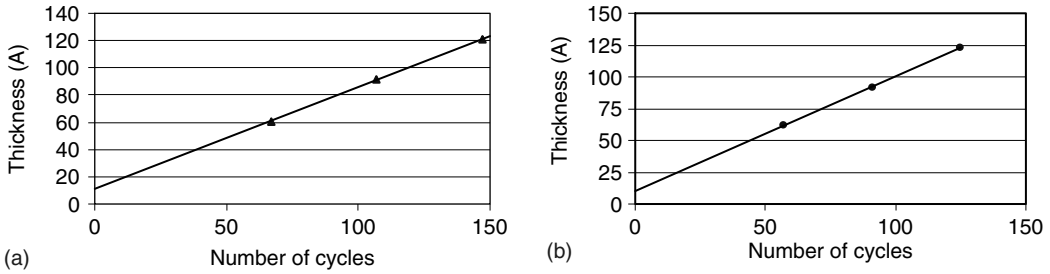


FIGURE 14.8 (a, b) The linearity of the deposition thickness with the number of exposure cycles for TEMAH/H_2O and TEMAH/O_3 processes. The offsets around 12 Å are due to inclusion of native oxide in the ellipsometric measurement. (After Liu, X., et al., *American Vacuum Soc., ALD-0*$_2$, Abstr. Book/CD, Seoul, Koera, 2002.)

Other variants of pulsed CVD can also lead to digital thickness control. Another approach uses a steady flow of one precursor and pulses the other. Both these pulsed CVD modes provide deposition with digital thickness control by controlling the dose of the pulsed precursor. One such process has been called atomic vapor deposition (AVD) [60] and typically provides approximately 100 Å/min deposition rates. See Figure 14.9 for a comparison of CVD, AVD, and ALD.

14.3.8 Initiation Processes

One of the most important topics for ALD technology is the initiation process. This topic includes surface preparation, chemically activated surfaces, and the conditions for initial continuous layer-by-layer growth. There are possibly two fundamental behaviors: (1) the surface is prepared with a dense reactive species (e.g., hydroxylation) that acts as a continuous template for the continuous ALD layer-by-layer deposition taking place from the first several cycles, and (2) island growth and coalescence occurs as the mechanism for overcoming nucleation delay and the transition to a constant growth thickness/cycle. The lecture notes of Steve George are useful for surveying the body of work carried out in this area [61]. Theoretical assessments by Puurenen are also providing insights [62].

The preparation of the surface for continuous layer-by-layer film growth is one of the most important technical challenges of ALD technology. It is desirable to start with a substrate *X* and deposit a continuous layer of a film type *Y* (not the same as *X*). It is inherent in the nature of interfaces of two dissimilar materials that a transition from substrate *X* to film *Y* may be accompanied by strain and localization of parasitic chemical species, Ideally, the layer-by-layer thicknesses per cycle (Å/cycle) moving through the interfaces are not substantially reduced.

The importance of continuous interface growth lies in the ability to achieve the full value of high-*K* materials. The deposition on silicon or metal substrates should transition from the surface to the high-*K* value in the shortest possible distance, without the formation of parasitic oxides (such as SiO_2 on a silicon substrate) or form other materials that may degrade the intended high-*K* values of the capacitor or gate dielectric. Yet, this can and does happen.

Looking at several case studies develops the topic.

14.3.8.1 ALD Oxides on Silicon Prepared by "HF-Last" and Chemical Oxides

If a silicon surface is pretreated with a HF last process to remove the native oxide and ALD is attempted directly on that surface, there is a large delay in deposition or a large incubation effect. An HF-last Si surface pretreatment followed by $HfCl_4/H_2O$ growth is characterized by "parabolic-sub-linear" growth behavior up to >40 cycles before linear growth proceeds [63]. In other words the HF-last preparation (H and F terminated) requires a longer initiation than a surface with an oxide.

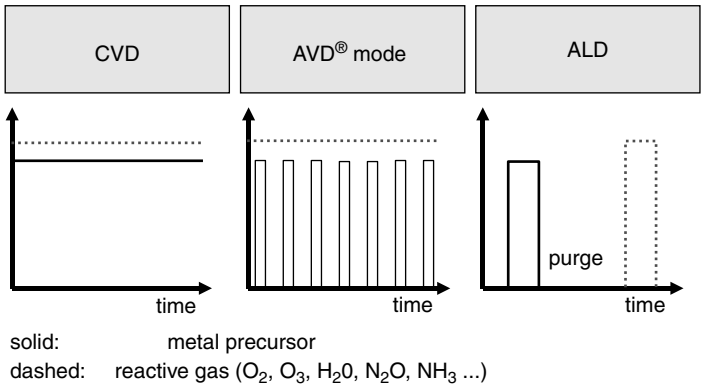


FIGURE 14.9 A schematic comparison of CVD, atomic vapor deposition (AVD), and ALD processes.

A phenomenological model uses the density of –OH terminations to obtain a nearly maximum growth rates from the first cycle. Surfaces were prepared with a variety of treatments including chemical oxide of approximately 7 Å, where growth linearity was extended down to approximately 4 cycles. On most surfaces the initial growth rates become linear once the film matrix becomes dense in –OH concentration or has substantial HfO_2. In other work, HfO_2 forms into nucleated islands as shown in TEM [64].

Si prepared with HF-last followed by a $ZrCl_4/H_2O$ ALD deposition sequence results in the formation of an interfacial oxide layer. After 40 cycles of $ZrCl_4/H_2O$ deposition there is 3 Å of interfacial SiO_2. After 24 days ambient storage there is 6 Å of SiO_2, with nominally 40A ZrO_2 on top of the SiO_2 immediately after deposition and after 24 days [65].

Si prepared with HF-last followed by TMA/H_2O, shows linear growth after about 10–15 cycles, which is the lowest incubation period among oxides grown directly on HF-last Si [65]. Aluminum from TMA is believed to directly bond to Si and Si–H. The bonding at the Si surface as may a rationale for the onset of relatively rapid continuous growth [66].

14.3.8.2 Initiation of Metals or Metal Nitrides on Oxides and Polymers

In the attempt to deposit metals on SiO_2, cluster islands are formed at high temperatures, but continuous conformal films are formed at low temperatures [12,67]. After 15 cycles, W grows linearly on SiO_2 [67]. W grows linearly from the outset on Al_2O_3 at 1 Å/cycle and after three cycles, deposits at approximately 2 Å/cycle. Al_2O_3 also grows continuously on W, as nanolaminates of W/Al_2O_3 have been demonstrated [67]. The Metals on low dielectric constant material (low-K) LoK process may be critical for future interconnect integration. The deposition of TiN directly on polymers is difficult. Al_2O_3 has been deposited on a low-K polymer followed by the immediate deposition of TiN. TiN using ($TiCl_4/NH_3$) forms continuous pin hole free layers after only 10 cycles of Al_2O_3 (TMA/H_2O) layers [68].

TiN grows essentially continuously on Al_2O_3 after several cycles. TiN is also a good initiation layer for other metal nitrides [69].

14.3.8.3 The Growth of Oxides at Nanolaminate Interfaces

An example is the delayed growth for the case of Al_2O_3 on a HfO_2 surface as part of a $HfO_2/Al_2O_3/HfO_2$ nanolaminate for gate evaluation. The chemistry was (TMA)/($HfCl_4$) with H_2O. Although there is a delay, it is not problematic. 6 Å of Al_2O_3 was placed on HfO_2 and followed by ALD HfO_2 layering [70]. About 1.5 times the usual number of cycles was used to obtain the desired Al_2O_3 thickness. HfO_2/Al_2O_3 nanolaminates were also reported using TMA chemistry, followed by Hf halide chemistry [71]. If the chemistries for the two metal oxides are the same (e.g., both halides), there may be a little delay in transitioning from one oxide to the other.

14.3.9 Nanolaminates

One of the earliest known publications of ALD nanolaminates was the layering of Al_2O_3 and TiO_2 using halide–hydride chemistry [8]. A recent example of the multiple nanolaminate with alloy layers is shown below in Figure 14.13 [69]. Nanolaminates may be developed for engineering or optimizing dielectric and leakage properties of various applications. For example: memory storage capacitors (71, J-H Lee-02), charge storage [72,73]; voltage coefficient of capacitance [74] or optical reflectivity [75] to mention a few.

- *Oxide nanolaminates*. Some nanolaminates are reviewed here in a catalog fashion with a short discussion with regard to application.
  - *Al_2O_3–HfO_2*. These nanolaminates are of interest for DRAM capacitor dielectrics [69,71].
  - *Al_2O_3–ZrO_2*. These nanolaminates were of early interest for gate dielectrics [65,76,77].
  - *Al_2O_3–Ta_2O_5*. "Aluminum–Tantalum–Oxide" nanolaminates has been of interest for potential DRAM dielectrics [78,79].
  - *Al_2O_3–TiO_2*. Another combination of potential interest for DRAMs [69,80].
  - *HfO_2–Ta_2O_5*. Still another combination for DRAMs [81].

Still other metal oxide combinations may be found in the literature and include: Al_2O_3–NbO_2; HfO_2–ZrO_2:HfO_2–Ta_2O_5; ZrO_2–Ta_2O_5:ZrO_2–Y_2O_3:NbO_2–Ta_2O_5:NbO_2–ZrO_2.

- Metal nanolaminates
  - *TiN–TaN*. The alloys of TiN and TaN have been mapped for application to barriers [82], see also the review by Kim [12].
  - *TiN–AlN*. The alloys of TiN and AlN have been characterized for possible use in higher thermal stable applications for DRAM electrodes [83].

Chemistries may (or may not) use a common family. There are chemical sequential compatibilities (or incompatibilities) in moving from one material to the other. For example, in the deposition of Al_2O_3/TiO_2 nanolaminates [8] a common halide–hydride chemistry was used. The terminating layer of Al–OH made by $AlCl_3$ chemistry is sequenced with $TiCl_4$, a similiar chemical reaction proceeds again producing HCl by-product. If chemistries are of a different family when moving between different film types of the compounds in the nanolaminate, there may be an incubation period associated with the transition between the different materials.

14.3.10 Rapid ALD by Limited Reactions

Although it may be surprising, limiting ALD exposures to values less than full saturation can increase the FDR (Å/min). This is because the CT can be reduced by a larger factor that the decrease in the ALD deposition rate (Å/cycle). In this approach, there may be a question of film quality and conformality, but in certain cases, these films still exhibit good properties as described at the end of this section [84].

Conventional wisdom has often used "overdosed precursor" ALD conditions to achieve complete saturation of the reactions [1,2,8]. The approach described here is a departure from that idea. "Rapid ALD" (RAD) may use a "Limited Optimized Reactions by ALD" (LORA) [84]. An optimization of the FDR (Å/min) occurs as a result of the maximum function of the product of increasing Langmuirian ALD deposition rate (Å/cycle) and the decreasing reciprocal CT (cycle/unit time).

FDR (thickness/unit time) approximately (Å/cycle) × (cycle/unit time)

$$\sim[1-\exp(-t_{ex1}/\tau_1)][1-\exp(-t_{ex2}/t_2)]/[\text{cycle time}] \tag{14.32}$$

where the first exponential function describes the saturation of one of the saturating precursor with time constant τ_1. The second function typically describes the slower saturating precursor, with time constant τ_2. The CT is: $CT = t_{ex1} + t_{r1} + t_{ex2} + t_{r2}$. Under the conditions that the first function is saturated and has a value K (constant in time), the FDR can be expressed as

$$\sim K[1-\exp(-t_{ex2}/\tau_2)]/[t_{ex1}+t_{r1}+t_{ex2}+t_{r2}]. \quad (14.33)$$

where t_{ex2} is the exposure time of the slower half reactions and τ_2 is its Langmurian time constant. The value K is related to the controlling value of the dose using the slower of the two half reactions. For example, if all time parameters are held constant except t_{ex2}, which is treated as a variable, then the FDR goes through a maximum because the exponential is a rising function and the reciprocal CT is a decreasing function with increasing t_{ex2}.

A case study for the maximization of the FDR is developed for TMA/H_2O chemistry. Al_2O_3 growth rates (Å/minute) using TMA and H_2O are plotted as a function of the H_2O exposure time. The results illustrating a limited reaction process are shown in Figure 14.10. A maximum is observed as the dose is reduced. The CT is about 0.5 s near the optimum FDR.

The FDR as a function of H_2O exposure time of the reactants exhibits a maximum, with a magnitude approximately 10 times higher that conventional ALD. An additional benefit of the LORA approach is that a dose used is less than the saturated value of at least one of the reactants, resulting in relatively high reaction efficiency; and since less reactants are used, less unused reactants have to be removed, resulting is a synergistic reduction in purge times. In this higher productivity mode, dose controlled exposures are used.

For the case of TMA/H_2O chemistry, Al_2O_3 LORA provides useful conformality, stoichiomentry, electrical properties, etc. Nearly 100% conformality has been demonstrated for aspect ratios approximately 40:1. The composition is substantially stoichiometric Al_2O_3 rutherford backscattering spectrometry (RBS). The electrical properties show breakdown fields greater than 8 MV/cm and low leakages are obtained. A film thickness uniformity has been achieved at the approximately 1% level for 100–2000 Å depositions. Densification may be appropriate after the deposition of such films [84].

14.4 ALD System Technology

14.4.1 General Description

Single wafer and batch designs have been implemented for ALD [1,9,85].

- *First Generation* ALD systems were introduced between 1975 and 1998. They were mainly characterized by horizontal flow and while the earliest trials used reactant removal by evacuation, even early systems used flow type reactors. Some were tube type designs, not unlike CVD horizontal tubes of the same era. Metal precursors other than TMA for aluminum were often solid halide precursors, proprietary gas valves were used [9] and pneumatic valves, when used, had at best only

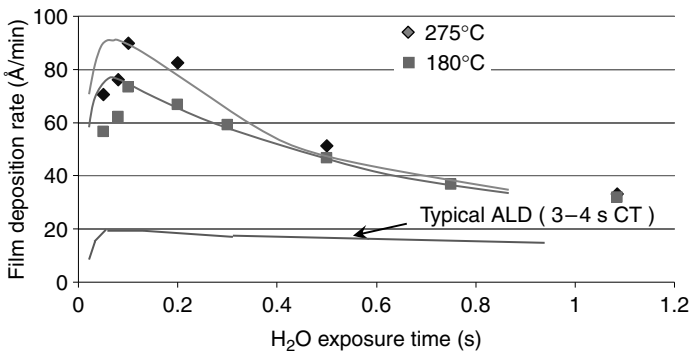


FIGURE 14.10 An illustration of enhanced deposition rates for limited optimized reaction by ALD. The data were obtained with TMA/H_2O and all parameters are constant except the H_2O exposure times. (After Kim, G. Y., et al., *ALD-04 Abstr. Book/CD*, Helsinki, August 2004.)

1 M cycle reliability. Some early batch, non-standard non-ULSI fab interface systems were commercialized for non-semiconductor and semiconductor R&D use.

- *Second Generation* ALD Systems were those introduced between 1998 and present times. Existing CVD wafer handling platforms that already had standard ultra large scale integration (ULSI) factory interfaces and wafer handling were modified for ALD operation [69]. For single wafer, axisymmetric vertical flow became common among many equipment suppliers, batch furnace type and multi-single wafer systems made appearances. For the most part, a shift to liquid precursors from metal precursors has been adapted. Pneumatic valves have improved in both speed of response and reliability. The number of cycles moved from 1 to 50 million mean time between failure (MTBF) level, now commercial components are typically used. Commercialization has had a good start [13], but in many cases still is limited by throughput (TP).
- *Third Generation* ALD systems are some systems introduced between the present times and going forward. Within the silicon semiconductor application area, these are for 300 mm using higher productivity processes. Improvements include advanced factory automation interface control and more mature systems from a maintenance point of view.

Whether single wafer or batch, ALD system description may include key subsystem features:

1. Chemical precursor sources with rapid, well controlled pulse delivery.
2. Delivery manifold(s) consisting of conduits, fast switching valves, and a gas distribution module.
3. Reactor vacuum chamber with a heated susceptor with controlled flow and pressure.
4. Exhaust conduits, downstream valves and pumps, and effluent management and control.
5. Real-time control system with approximately 10 ms control accuracy and precision.
6. Industry standard wafer delivery capabilities.
7. ULSI factory automation.

Precursor source technology is of major importance for the success of future ALD. Although many precursors and chemistries have been demonstrated [45], few, so far are actually suitable for large-scale commercial use. Liquid precursors of higher vapor pressure, (so far limited to TMA and titanium tetrachloride…) in the approximately 10 Torr region and higher thermal stability against decomposition at high temperatures are needed. Synthesis of new chemical precursors for metals like Hf, Zr La, Ta, and Mo is an ongoing activity, and the development of hafnium alkyl amides has been an important advance.

A short list of the general requirements for useful precursors for ALD includes:

a. High reactivity with –OH or –NH terminated surfaces.
b. High reactivity with each other (metal precursor and non-metal precursor).
c. Atomic layer deposition saturating reaction chemistry.
d. Volatile reaction by-products that do not react with the depositing film.
e. High volatility.
f. High thermal stability.
g. Low retention of C and H in the deposited film.
h. Low cost, and cost per wafer for the deposited thickness.

Various designs of a delivery manifold consisting of conduits and fast switching valves can be used. Important factors are the speed, repeatability, and reliability of fast switching valves used to pulse in the precursors. “Chemical Dosage” that is, the precursor partial pressure for a certain exposure time, is typically required to be approximately 100 mTorr for 0.1 s for single wafer systems, but may be (necessarily) longer for batch systems. In any case, high partial pressure of the precursor is important. The valves themselves today have actuation speeds of 10 ms, with repeatability of 1 ms with 50 million-cycle reliability [86].

In the (historical) conventional wisdom view of ALD, the delivery of precursor is “unimportant” as long as the dose is above the saturation timing edge. Once saturated, there is no need to deliver precursor uniformly, or alternately, there is no penalty for non-uniform distribution during delivery. However, as

many real ALD processes have saturations that are soft, and as it is not preferable to waste precursor, it is desirable to operate near the edge or earlier into the kinetic regime. Under this mode, gas distribution and dose control becomes important. The residence time of the precursor moving through each volume in the ALD system is significant, and movement through the gas distribution module is no exception.

Minimal reactor volumes (e.g., limited by the "reaction" space above the wafer surface), minimal reactor parasitic surface areas and avoidance of re-circulations are factors in good reactor design. The direction of precursor flow relative to the substrate has been used in either horizontal [1] or vertical flow configurations. See Figure 14.11 and Figure 14.12. The vertical flow arrangement may be preferred for smaller footprint, and ease of gas mass transport engineering, which favorably provides for effectiveness of the chemical removal and the management of minimal symmetric parasitic CVD over the wafer radius rather than the wafer diameter.

Pumps with high capacity are important to achieve rapid chemical removal. The trapping of downstream effluent is another consideration as the reactants are, by design, highly reactive and will invariably, unless selectively diverted, form heterogeneous wall deposits down stream.

The design and operation of a flexible vacuum system is important. Generically, we refer to this as time-phased multi-level residence-time (TMR) providing a variable residence time approach for ALD. The residence time may be varied during different periods of the ALD cycle. Use of low residence time (PV/F) with, e.g., relatively *high flow*, F reduces the removal time of the by-product and unused reactant. In principle, either pressure (P) or flow (F) may be varied during the cycle, or both may be varied to affect the desired change in residence time. The method of running different flow conditions during ALD exposure and removal periods has been practiced in the 1990s [87]. A particular technique to implement different flows is described as synchronously modulated flow

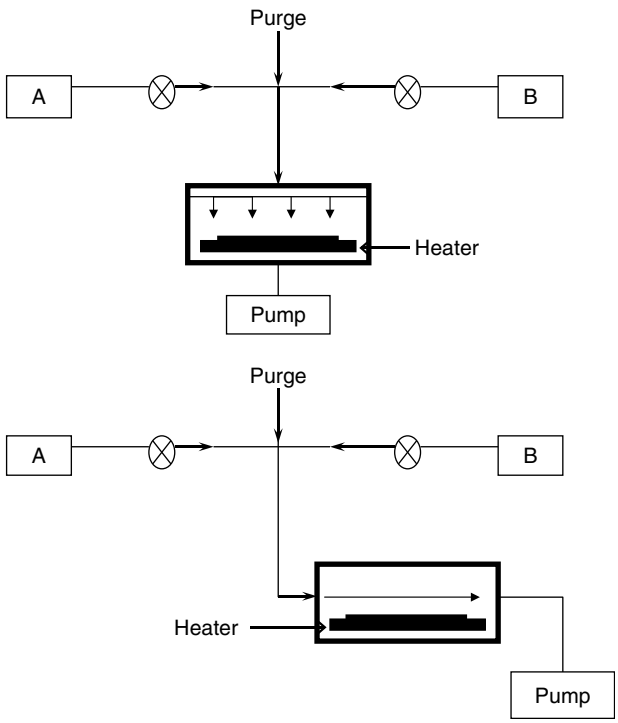


FIGURE 14.11 Schematic of single wafer reactors with generic source for A and B chemicals showing vertical and horizontal reactor flows.

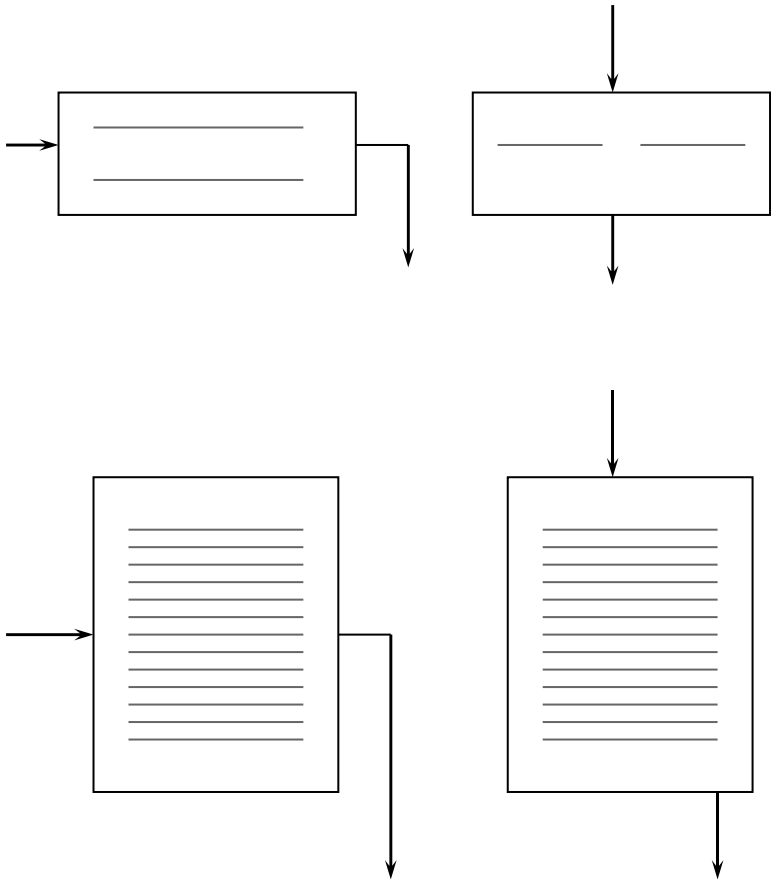


FIGURE 14.12 Schematic of batch reactors with "multiple single water" and vertical batch configurations. Various flow configurations are also shown.

draw (SMFD) [88]. The benefits of variable residence time are: *relatively high residence time during exposure* increases the efficiency of precursor use and *relatively low residence time during purge* decreases the removal/purge period.

14.4.2 ALD Systems

14.4.2.1 Single Wafer Systems

Reactors for single wafer ALD operation are common. A single wafer CVD-type reactor, suitably modified for short residence time and customized with ALD software command control, and manifold hardware permits ALD operation [69]. Such systems can be ULSI compatible at the outset and have cluster tool architecture. The operation of such a system is, in many ways, common, and transparent to CVD operation with similar architecture, but in other ways is not important. The flexible software controls and manifold allow the deposition of nanolaminates. The nanolaminates allow for engineering of compositional control, and by using a cluster tool architecture films deposited in metal and dielectric modules can be interfaced. An example of early implementation of dielectric and metal modules on the same cluster hub is the Genus *LYNX*-ALD or StrataGem^{300}. Single wafer systems are especially amenable to plasma-assisted operation. When operated with axisymmetric vertically injected chemical reactants,

they have minimal parasitic CVD effects. For a given minimal removal (purge) period, the parasitic CVD is engineered over the radius of the wafer and not the diameter. Single wafer tools also have the capability with suitable modular chemical sources to extend to multiple technology generations. The capital cost can be higher than batch systems, yet the chemical usage and costs to operate can be lower and can be operated without backside deposition. With process enhancements such as LORA or TMF, the single wafer reactor use can be extended to thicker films.

Single wafer systems features

- Process module clustering.
- Multi-nanolaminate insitu deposition within preventive maintenance (PM) and within PM-to-PM cluster.
- Limited Optimized Reactions by ALD and TMR enhancements.
- Backside deposition can be prevented.
- Vertical inject (radius control) or horizontal inject (diameter control).
- Radio frequency (RF)-Plasma direct or remote deposition or in situ clean possible.
- Lower and optimized precursor use.
- Thinner film advantageous, may be TP limited for thicker films.

In summary, the leading attributes of single wafer systems are: its flexibility (fully capable of nanolaminate and interfacing different films) as well as plasma operation, extendability, and maintainability.

Suitably designed ALD systems [69] can be programed to deposit one type of film and then another such as Al_2O_3/HfO_2 laminates or a mixture of two of more elements, such as $(Al_2O_3)_x\,(HfO_2)_{1-x}$, which can be made as one of the distinct layers in the nanolaminate. Such ALD flexibility is shown in Figure 14.13, by the high resolution TEM. Typically as-deposited nanolaminates are well defined with nearly atomically sharp interfaces. In certain cases, the actual smoothness will depend on the thermal history of the nanolaminate. Nanolaminate films have been deposited between 200 and 450°C and a small increase in roughness is observed at higher temperature [40]. This is possibly due to different initiation conditions and/or in situ grain growth during film deposition. It is also known that under high temperature post-deposition anneals (PDA) of 950°C that distinct layered nanolaminate films interdiffuse [39].

14.4.2.2 Batch Systems

Minibatch (25 wafer) with a cluster architecture and batch (50 wafer and/or greater) using vertical furnace configurations have been operated in the ALD mode. These systems may have relatively lower capital cost. They require provision for backside deposition, and engineering for limiting parasitic CVD over the diameter of the wafer. Plasma operation and insitu nanolaminate or alloy operation is not convenient, remote plasma has been introduced. Maintenance requires tube change out as in thermal furnaces and there may be added cost and management of high precursor use. Extendibility is challenged considering there is an increasing degree of difficulty of bringing higher doses to the batch of wafers in a shorter period of time.

Batch reactor features

- Clustering approach possible, requires batch handling.
- Some nanolaminates are successful, but some may be difficult to develop.
- Lower capital system cost.
- Remote plasma approach is possible, but with difficulty.
- Advantageous for thicker film, but may be load time limited for thinner films.

In summary, a wide variety of ALD reactor architectures are available for the deposition of ALD.

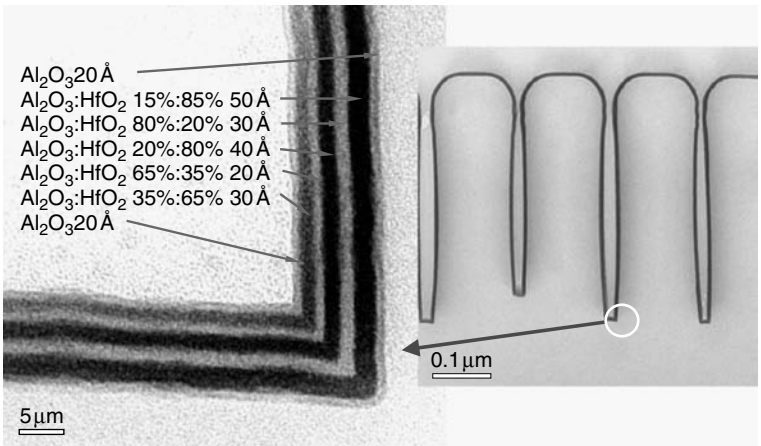


FIGURE 14.13 Demonstration of conformal coating in test trenches with nanolaminates of Al_2O_3 and different alloys of Al_2O_3 and HfO_2. The high-resolution TEM on the left is a typical lower right corner of the cross sections. (After Sneh, O., et al., *Thin Solid Films*, 402, (2002): 248.)

14.5 Applications

14.5.1 Higher-*k* Oxide Capacitors on Chip

Referencing Table 14.4, note that the difference in aspect ratios for cylinder/stack and trench capacitors (19 vs. 55 in 2004). The present trench capacitor architecture has more active area than the stack/cylinder type in the near term years. In 2010, the percentage difference in aspect ratio is narrowed (72 vs. 94). As a result, in the near term trench technology can delay the switch to high-K and MIM by one generation. Al_2O_3 ALD is in production using cylinder/stack architectures. In 2007 and beyond, other high-K's are needed (indicated by "+", later roadmaps show HfAlO, ZrAlO, and LaO-based materials). They may be, however, HfO_2 or ZrO_2, and these may be alloyed or nanolaminated with Al_2O_3 [39]. In the migration from SIS to metal-insulator-semiconductor (MIS) to MIM, both stack and trench architectures will need substantially conformal metal electrodes, and this is an opportunity for ALD. The lower electrode will have higher conformality requirements than the upper metal electrode. ALD TiN films are already developed for DRAM bottom and top electrodes, beyond 45 nm Ru is being considered [90].

TiN is used for DRAM stack capacitor electrodes [39] and deep trench upper electrodes [92]. TiN has good thermal stability allowing deep trench process integration, while maintaining low resistivity through 1050°C. The expectation is that ALD will continue to be required for the high aspect topology for capacitor metal electrodes.

TABLE 14.4 Capacitor

Year of Production	2004	2006	2007	2009	2010
Structure/Cyl	MIS		MIM		MIM
Structure/Tr	SIS		MIS		MIS
DRAM ½ pitch (nm)	90	70	65	50	45
DRAM product	1G	2G	2G	4G	4G
Aspect ratio/Cyl	19		26		72
Aspect ratio/Tr	55		78		94
Diel/Cyl	Al, TaO		Al, TaO+		Al, TaO+
Diel/Tr	SiON/Al_2O_3		high-K		high-K

The active area of the capacitor is much larger than the planar area of the silicon wafer. A 300 mm wafer has an active device surface area of 700 cm^2. The active area of a DRAM in 2004 is 4000 cm^2; and in 2010, it is projected to be 16,000 cm^2. This large active area-density places a burden on the rate of chemical precursor delivery to the surface of the wafer if productivity is to be acceptable. The capacitor high-aspect ratio and high area are a challenging requirement, but is within the capabilities of ALD.

Gordon et al. has described the model for 100% step coverage in extreme high aspect ratio structures [89]. A key overarching assumption is: reactive molecules used to coat the inside of the hole are limited by the *diffusive flow*. This takes a much longer time than reactions on a flat surface.

The key elements of the Gordon model are described. The amount of reactant needed to cover the entire surface per cycle is the value of the *saturation surface density S* (number of atoms/cm^2) $\times$ the entire surface area per cycle. The time required to supply saturation dosage S is approximated by $t = S/J$, where J is the flux. An isotropic flux J of reactant molecules approaching the hole is $= P/(2\pi\, mkT)^{1/2}$, where P is the partial pressure of the precursor, m is the precursor's molecular mass, k is Boltzman constant and T is the temperature.

The flux is assumed constant and there is no depletion of precursor partial pressure during the exposure time, t. Then the exposure dose for saturation on planar surfaces is:

$$\text{Pt} = S(2\pi\, mkT)^{1/2} \tag{14.34}$$

where Pt is the "exposure," expressed in units of Torr-s. To develop an expression for the exposure into a high aspect hole, there are two assumptions.

1. There is molecular diffusive flow inside the hole. The condition for molecular flow is that the Knudsen's number (Kn = mfp/d) be > 1, where mfp is the mean free path of the molecules and d is the characteristic dimension of the hole. At 200 mTorr pressure and 200°C temperature, the mfp is in the mm range and the molecular flow assumption is confirmed.
2. The vapor by-products do not redeposit or etch the deposited film.

To these we add another consideration: the reactants must be substantially removed from the hole before the next reactant's half reaction occurs, avoiding CVD bread loaf coatings near the top of the hole; the conditions are selected to avoid gas phase reactions in the hole.

At a distance λ down the hole shown in Figure 14.14, J is reduced by the clausing factor [89] according to $J = J_s/\{1 + (3\lambda p)/(16A_{\text{hole}})\}$, where J_s is the flux at the orifice of the hole, p is the perimeter of the hole and A_{hole} the cross sectional area. Inserting the value of J and integrating the time to reach a hole depth L, the exposure needed to cover the sidewall is $(\text{Pt})_{\text{sidewall}} = S\ (2\pi\, mkT)^{1/2}\ [4a + 3/2a^2]$, where d by $a = L_p/4A_{\text{hole}}$. The exposure needed to coat the bottom of the hole, $S(2\pi\, mkT)^{1/2}\ [1 + 3/4a]$, needs to be added to that of the walls.

The total exposure needed to coat the sides and bottom of the hole completely is,

$$(\text{Pt})\text{total} = S(2\pi\, mkT)^{1/2}[1 + 19/4a + 3/2a^2], \tag{14.35}$$

This equation is plotted in Figure 14.14. The model has been experimentally verified using doses from lower than that required to coat the trench to more than that required [89].

High aspect ratio features in DRAM and other semiconductor devices have been coated using ALD. Figure 14.15 shows the coverage of the high aspect ratio deep trench capacitors by ALD. The structure has a bottle (re-entrant) architecture and 50:1 aspect ratio. A 45 Å Al_2O_3 film conformally coats the trenches. The dielectric quality meets conformality, electrical film reliability, and uniformity over wafer diameter requirements [91].

In Figure 14.16, the stack capacitor is illustrated for conformal coverage with nanolaminates of Al_2O_3 and HfO_2 [39]. The use of liquid precursor TEMAH for Hf allows better conformal deposition relative to $HfCl_4$, which is a solid with low-vapor pressure. HfO_2 film's electrical characteristics are good.

This class of capacitor applications includes DRAM and eDRAM below 100 nm [39,91], RF-decoupling capacitors with thicker films and metal electrodes [92] leading ultimately to higher performance MIM

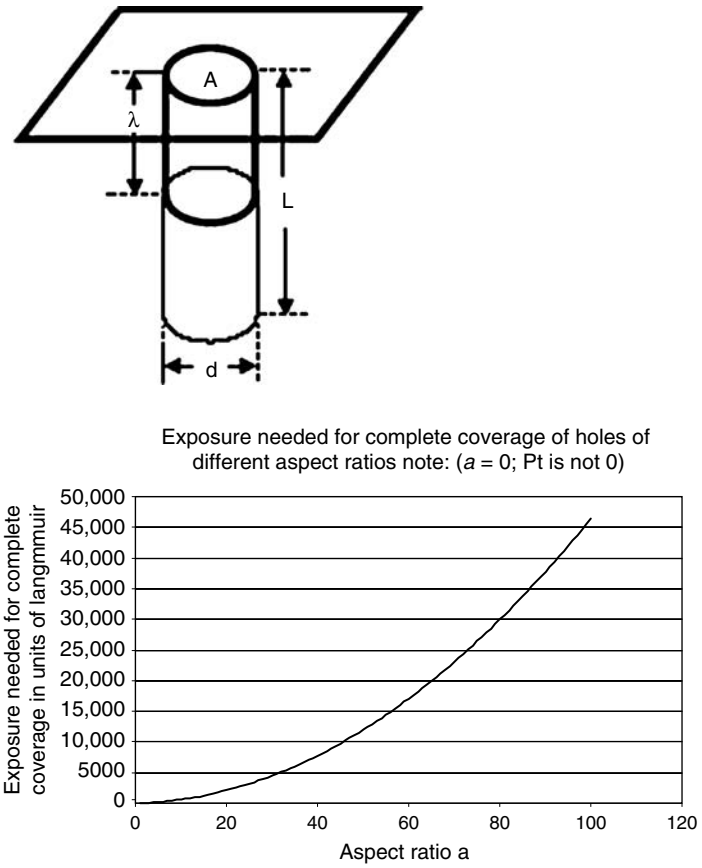


FIGURE 14.14 Cross section of a hole modeled for step coverage. The graph shows the exposure required to obtain conformal ALD coating increases as the square of the hole's aspect ratio. (After Gordon, R., et al., *Chem. Vapor Deposition*, 9, (2004): 73.)

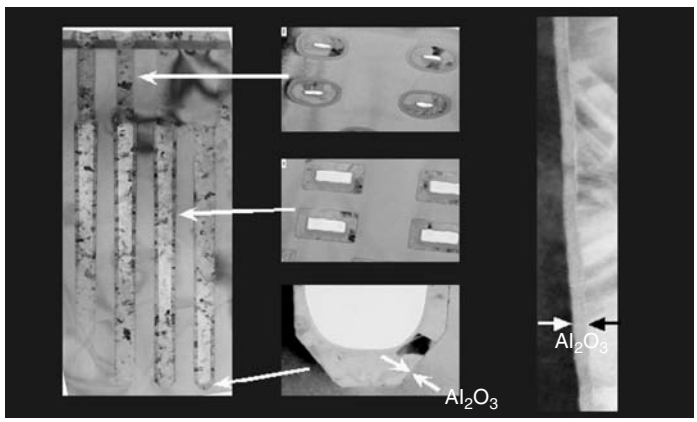


FIGURE 14.15 Demonstration of conformal coating in deep trench dynamic random access memory (DRAM) device structures using Al_2O_3. The trench structure is a re-entrant "bottle" design. The high-resolution TEM in the center shows conformal coatings across the etched facets at the bottom of the trench. (After Gutsche, M., et al., *IEDM Tech. Dig.*, (2001): 411.)

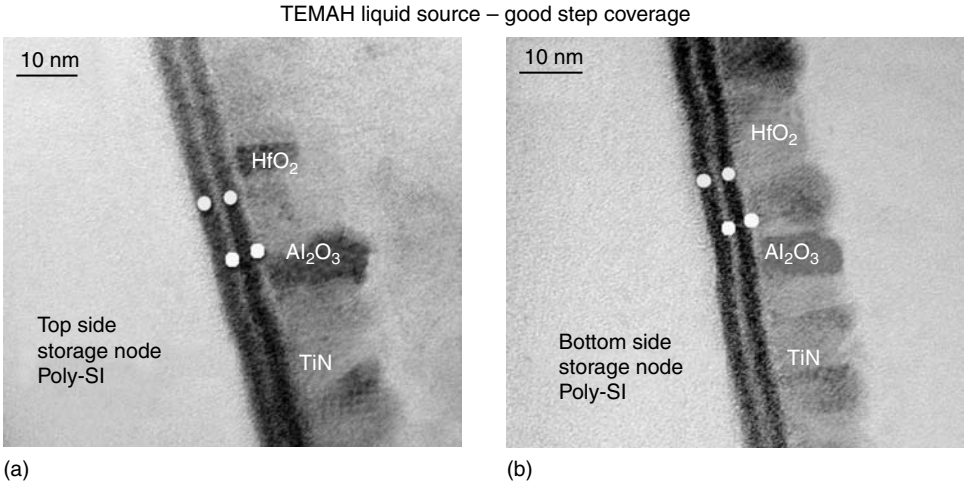


FIGURE 14.16 Demonstration of conformal coating in DRAM capacitor stacks using Al_2O_3 and HfO_2 nanolaminates. The high-resolution TEM shows conformal coatings of the stack element. (After Lee, J. H., et al., *IEDM Tech. Dig.*, 221, 2002.)

structures. The DRAM capacitor application is driven by the need to increase capacitance density under scaling. It is projected in the ITRS that DRAM capacitor deep trenches made circa 2010 will require much higher film surface area than that used today. Even the DRAM's produced today have $\sim 10\times$ the active areas of planar surfaces. Trench architectures will require step coverage on aspect ratios approaching 100:1 at 45-nm feature size. At this point the active areas will be approximately 20 times the planar silicon area. The challenge to provide conformal films on such high-density structures can be met, if there are significant improvements and developments in chemical precursors, delivery systems, optimized processes, and ALD operating systems. Furthermore, multi-wafer systems are expected to bring productivity enhancements to the ALD tool set. All this assumes that properties such as EOT and leakage are met.

RF and decoupling capacitors require thicker films and, therefore, also require higher deposition rates. Initially implemented on planar or low aspect-ratio structures, eventually these capacitors will require conformal coatings on higher aspect-ratio capacitors. These capacitors have different electrical requirements than DRAM capacitors, notable they are challenged by the requirement of an engineered, low voltage coefficient of capacitance.

14.5.2 Advanced Dielectrics and Metal Gates

Referencing the elements in Table 14.5, CMOS gates in production today are planar and migrating to fully depleted SOI. The development of advanced dielectrics may be approached with a variety of processes such as CVD [93] or even PVD [94]. Some device developers are considering CVD to be the technology of choice. Recently, ALD gate performance has achieved parity results with CVD [3]. The comparisons are with dual-doped polysilicon. These high-K solutions are particularly promising for low power use, since the initial EOT requirements appear to be nominally met. Having said that, the challenges to obtain controlled-EOT less than one nm for high performance (high drive current) beyond 2007 are not yet obtained. The best technology (e.g., ALD or CVD), which enables high performance for less than 1 nm EOT, will be considered.

Atomic layer deposition has demonstrated a core capability to provide uniform film properties for advanced gate application. Figure 14.17 shows the good distribution of gate leakage currents for N-channel metal-oxide semiconductor (NMOS device) using 30 and 40 Å ALD HfO_2 vs. 21 Å SiO_2 control, with a pre-deposition NH_3 surface treatment [70].

TABLE 14.5 Gate

Year of Production	2004	2006	2007	2009	2010
Structure	Bulk Planar CMOS			FD–SOI	
EOT Lo power (nm)	1.5	1.3	1.2	1	0.9
EOT MPU Hi perf (nm)	1.2	1	0.9	0.8	0.7
EOT interface Ox (nm)	0.7	0.5	0.5	0.4	0.4
EOT high-K (nm)	0.5	0.5	0.4	0.4	0.3
Phys Thk high-K (nm)	2	2	1.6	1.6	1.2
Silicide thick (nm)	20	19	17	14	13
Elec contact thick (nm)	20	15.4	13.8	12	10.8

Figure 14.18 shows the results of Jung et al. illustrating that HfAlON ALD films were enabled by PDAs using NH_3 to provide approximately 80% mobility relative to SiON controls [95].

A summary of various films and deposition methods is shown in Figure 14.19, comparing HfAlO, HfAlON, and HfSiON. Deposition methods include CVD [93], PVD [94], and ALD [3] for HfSiON. The ALD results for HfAlO and HfAlON are compared as well [95,96].

Many gate development efforts have been carried out using ALD dielectric depositions. The current material of choice is HfSiON [3,93,94], and although this material is a complex quaternary, ALD has potential to control the composition of the appropriate elements. An equally important factor is the engineering of ultra-thin interfacial oxides less than 0.5 nm, otherwise future EOT goals may not be met. These requirements are shown in Table 14.5, for example 0.8 nm EOT for 2009. Assuming a metal gate with a zero depletion width, and the oxide is partitioned as 0.4 nm for SiON and 0.4 nm for high-K, then the 0.8 nm EOT value is achieved. If an effective High-*K* of 16 is assumed, the physical thickness of the high-K is approximately 1.6 nm. The tunnel current would be expected to be comparable to a 20-Å SiON film, but the EOT should be approximately 1/3 of the SiON value (8 Å instead of 20 Å.)

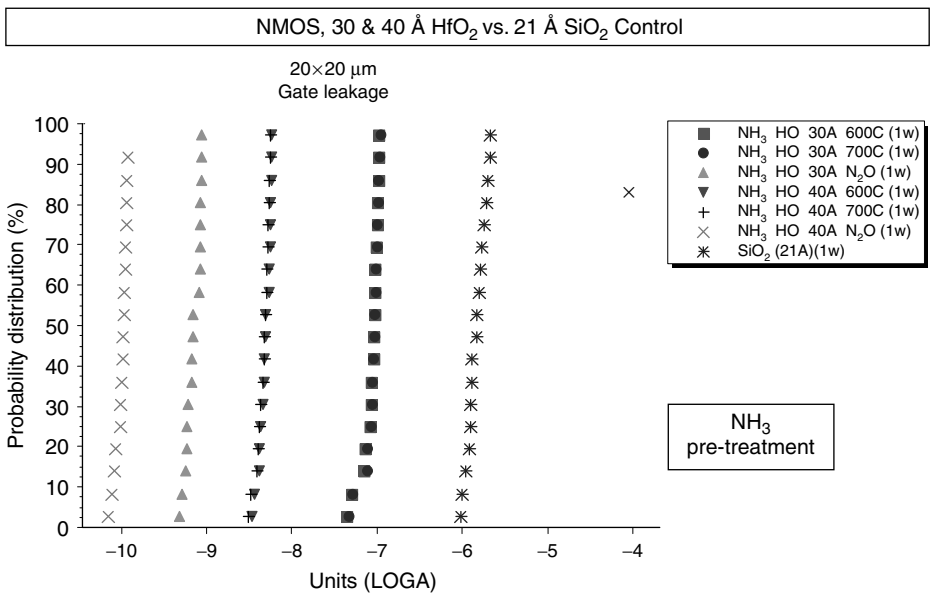


FIGURE 14.17 Leakage distributions for gate test structures showing tight distributions with lower leakage and equivalent oxide thickness (EOT) values than a 2 nm SiO_2 control. The surface was NH_3 pretreated; the EOT of HfO_2 films are as low as 1.4 nm. (After Londergan, A. R., et al., *The Electrochem. Soc. Proc.*, 2003.)

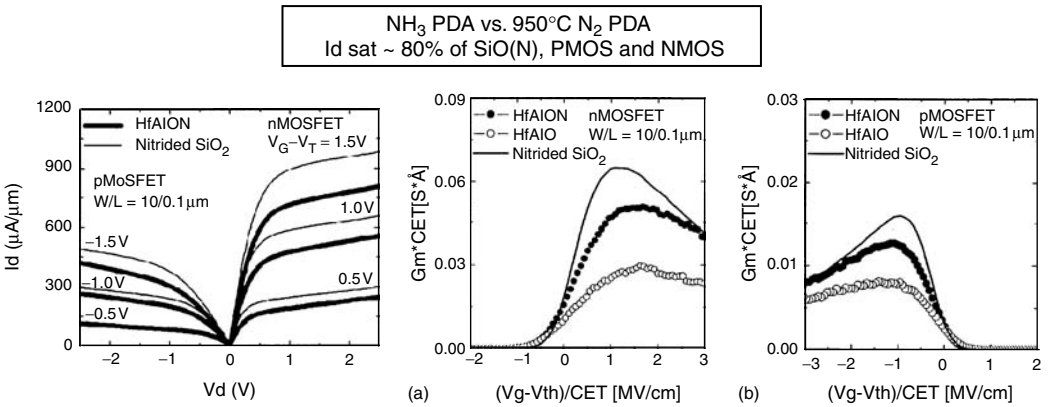


FIGURE 14.18 Drain currents for NMOS and PMOS devices using HfAlON layers, the HfAlO was by ALD and the nitrogen containing interface was formed by active post deposition annealing. G_m values are 80% of the nitrided SiO_2 control. (After Jung, H. S., et al., *IEDM Tech. Dig.*, 853, 2002.)

Implementation of high-K dielectrics with poly gates is limited by Fermi level pinning that shifts the threshold voltages to high values and shows channel mobility degradation [4]. As a result, solutions have moved to metal gates, which may alleviate the Fermi pinning and has shown promise for higher mobility. Metal gates also provide for a step function reduction in EOT because of a reduction of depletion in the poly gate. This assumes the metal gate's carrier concentration is suitably large.

ALD TiN [98,99], MoN [101] and pulsed CVD for TaN [60] are also considered for gate applications. ALD and pulsed CVD metal films provide precise thickness and composition control to set up suitable work functions for NMOS and PMOS devices.

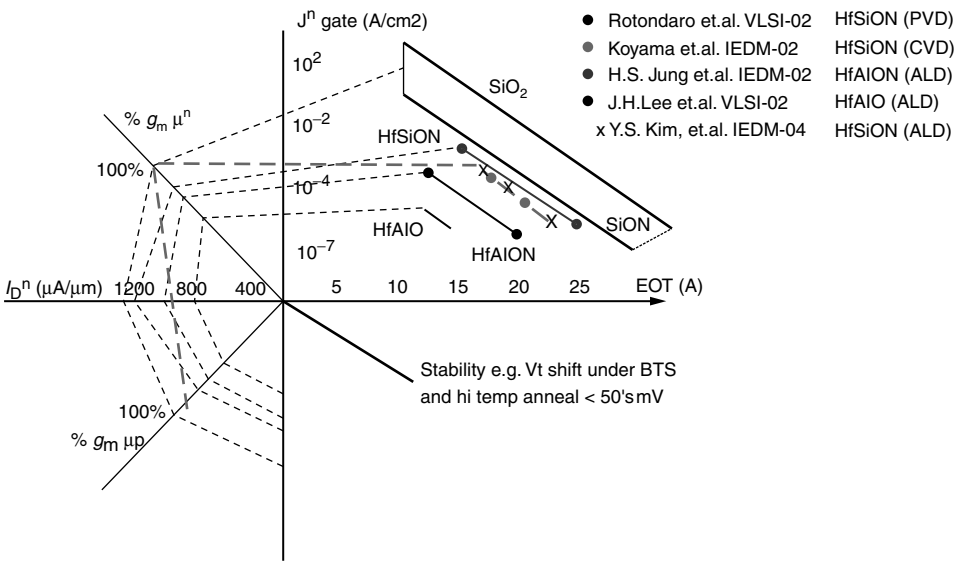


FIGURE 14.19 "Spider Chart," comparing various advanced high-K gate results. Although HfAlO has the best J-EOT performance, the HfSiO(N) has the best carrier mobilities, stability and integration potential. All results are for dual-doped poly gates.

The current failure to develop a high mobility bulk planar solution has led to parallel development of high mobility using strained silicon. It is now likely that high-K dielectrics will be developed in combination with strained Si and metal gates, especially as strained silicon comes into widespread production.

Beyond 2010, short channel effects drive the gate architectures to 3D structures such as vertical MOSFETS with FINFETs architecture. These structures will require dielectric and control gate electrode deposition on vertical walls, and as high-K dielectrics and metal gates are to be used, the technology of choice could be ALD.

In summary, since SiO_2 and SiON gates have reached the tunnel limit at thicknesses of 1–2 nm, there has been a significant initiative to develop thicker, higher-*K* materials with low gate leakage. After considerable efforts on ZrO_2 with poly silicon gates, the development community has turned to HfO_2/poly and then to HfAlO(N) and HfSiO(N), and also using metal gates for obtaining the lowest EOT. Different metal gates are being pursued for CMOS application. Initiatives in "Low Power" applications will be followed by "High Performance" applications, the latter including high mobility (higher speed) as well as low leakage capabilities. Today advanced gate stack films may be made with CVD, PVD or ALD [97].

14.5.2.1 Interconnect (Barriers and Cu Seeds)

14.5.2.1.1 The Need and Timing of ALD for Interconnects

Interconnect requirements are challenged by the control of barrier thickness in moderate aspect ratio features, and currently are addressed in this Handbook in greater detail in the Chapter 13 on CVD.

However, referencing Table 14.6 and combining via and wiring aspect ratios we have an AR about approximately 3.3 regardless of the generation. Aspect ratios in logic interconnect wiring levels are not scaling because of the adverse effect of capacitive cross talk between the intermediate wiring levels. ALD enables controlled conformal barrier thickness, allowing barriers to be scaled without concern of thickness variations along interconnect feature edges. The thickness control has the capability to improve the cross sectional area of the Cu wiring by reducing the barrier thickness. This illustrated in Figure 14.20 [98]. Other rationales for the use of ALD include stress and compositional control. Atomic layer deposition is in principle well suited to provide improvements in these areas.

DRAM contacts, on the other hand, have high-aspect ratio challenges that track the aspect ratios of the stack/cylinder DRAM capacitor. In 2010, according to the ITRS the DRAM contact aspect ratio is greater than 20 and this is an opportunity, for example, for ALD TiN barrier and ALD W or other metal-fill materials.

14.5.2.1.2 ALD Pathway Potentials in Interconnects

The review by Rossnagel [99] on ALD potential interconnect and contact applications is summarized here

A current *contact* process of record is: CVD TiN/ex-situ PDA/W nucleation/W fill.

In future, the resistivity of W may be limiting and reduced with a higher conducting material, if suitable for manufacturing. There is presently no-known ALD Cu process suitable for manufacture, and ALD Ru has been demonstrated about half the resistivity of W, but there is a cost of material issue with Ru. The ALD WN/W/CVD W has been reviewed as well [100].

TABLE 14.6 Interconnect

Year of Production	2004	2006	2007	2009	2010
Logic Structure	Dual Damascene ~8 Intermed Levels				
DRAM Structure	Cylinder Stack w Hi AR				
Wiring AR intermed lev	1.6	1.6	1.7	1.7	1.7
Via AR intermed lev	1.5	1.6	1.6	1.6	1.6
DRAM contact AR	15	16	16	17	>20
Logic bar thk (nm)	10	8	7	6	5

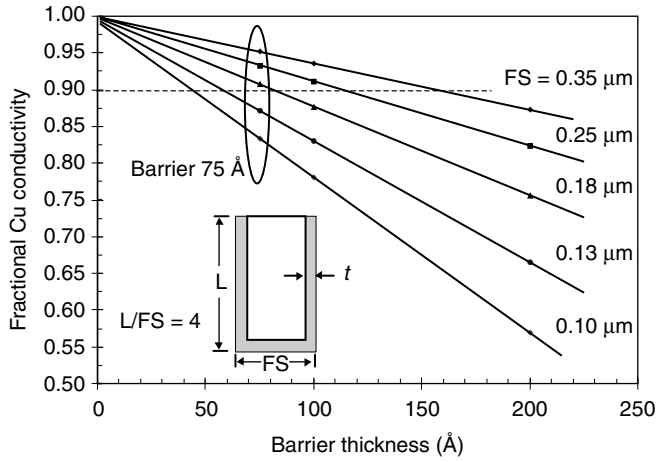


FIGURE 14.20 Fractional Cu conductivity as a fuction of the barrier thickness. The barrier's conductivity is assumed negligible relative to Cu. The trends with decreasing feature size are noted. After Ivanova et.al., Integration of PECVD Tungsten Nitride as a Barrier layer for Copper Metallization. *Mat. Res. Soc. Symp. Proc.* 564 (1999): 321.

A current *multi-level interconnect wiring* process is: ionized PVD TaN/etchback/i-PVD Ta/iPVD Cu.

This sequence uses four processes (iPVD TaN, etch, iPVD Ta, and iPVD Cu seed.). Thick Cu is electroplated on top of the iPVD Cu. Unfortunately, ALD TaN creates a high-resistivity phase Ta_3N_5 (approximately 1 ohm-cm level). Other ALD processes may be successful in creating lower resisitivity ALD Ta(N). A future (simplified) multi-level interconnect wiring process may be developed for the 45 nm node: ALD TaN, ALD Cu. This could be an elegant solution. But it assumes the development of low-resistivity ALD Ta(N) and ALD Cu.

Equally significant for interconnect area is the pore sealing for use with porous low-K materials. A modified (non-ideal) ALD that is designed to bridge the pores, followed by ideal ALD is needed.

14.5.2.1.3 Composite Engineered Barriers by ALD (CEBA)

Atomic layer deposition has the unique capability to form nanolaminates. In the interconnect area, this might be exploited to engineer the adhesion to the lower interface, to optimize the density of the barrier and set the texture of the upper surface of the barrier. This was illustrated in work using layered TiN/TiN–TaN alloys, where the surface composition is Ta-rich to provide crystallographic texture for Cu orientation [82]. The concept is shown in Figure 14.21.

14.6 Summary of Current Status and Outlook

- Many ALD processes are established and already applied. Capacitor applications are in production using dielectrics at 90 nm. Higher *K* dielectrics and metal electrodes are envisioned for use below 65 nm. Gates applications are limited by control of surface oxidation in relation to the EOT budget. Interconnect applications are anticipated below 55 nm.
- Equipment is in transition to the third generation
  - Productivity enhancements by process, vacuum engineering, overhead, and multi-wafer systems are currently being introduced.
- Future Vision
  - Both silicon and nano device applications with high topology are in the wings. There is a potential role of ALD as an enabling pathway to both the future silicon roadmap and broader nanotechnology implementations.

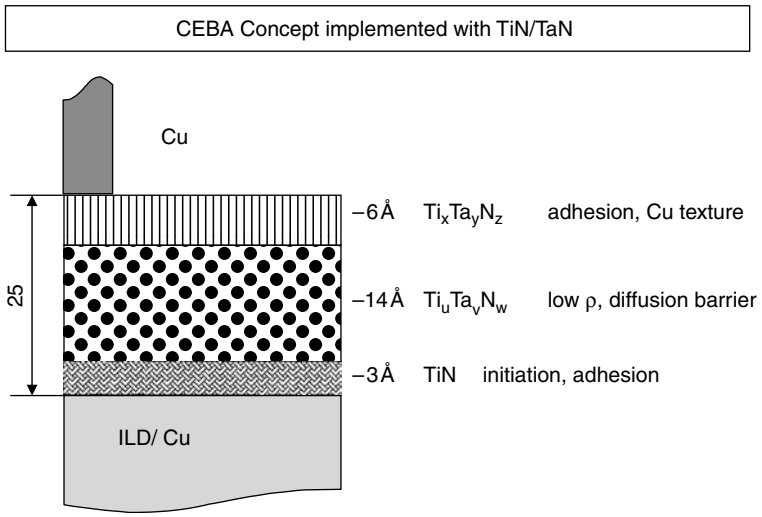


FIGURE 14.21 Schematic of composite engineered barriers by ALD (CEBA) showing the initiation layers, the barrier and the top layer providing adhesion and texture for the overlaying Cu layer. (After Londergan, et al., *Engineered Low Resistivity Titanium–Tantalum Nitride Flms by ALD*, MRS Spring Meeting, San Francisco, CA, 2001.)

The future outlook for ALD includes use for 3–5 compounds, potentially quantum wells, tunnel diodes, micro fuel cells, advanced phase shift masks, extreme ultra violet reflective masks, porous substrates, mono-dispersed spheres and all varieties of nanotechnologies with topology challenges.

Acknowledgments

The author thanks Jeremie Dalton, Carl Galewski, Steve George, Roy Gordon, Zia Karim, Gi Kim, Xinye Liu, Ana Londergan, Charles Musgrave, Sas Ramanathan, Mikko Ritala, Ofer Sneh, Anuranjan Srivastava, Gillian Zaharas, and Zhihong Zhang for their interest, discussion, and contributions.

References

1. Suntola, T. "Atomic Layer Epitaxy." In *Handbook of Crystal Growth*, Vol. 3, edited by D. T. J. Huerle, 601. Amsterdam: Elsevier, 1994, chap. 14.
2. The International Technology Roadmap for Semiconductors (ITRS), *Semiconductor Industry Association*, 2003 ed. SEMATECH: Austin TX, 2003, pp. 356–357 and see also 1st reference ALD in ITRS, 1999 ed. SEMATECH: Austin, TX, 1999, p. 76.
3. Kim, Y. S. et al. "Characteristics for $HfSiO_x$ Using New Si PRECURSORS for Gate Dielectric Applications." *IEDM Tech. Dig.* 20.4 (2004): 511.
4. Chau, R. et. al. In *Proceedings of AVS 5th International Conference on Microelectronics and Interfaces*, 3, 2004.
5. Kautsky, M. et al. *Applications of Atomic Layer Deposition for Magnetic Heads*. San Jose, CA: American Vacuum Society, 2003 (ALD-03).
6. Park, D. et al. "3-Dimensional Nano-CMOS Transistors to Overcome Scaling Limits," *7th ICSICT*, Beijing, A1.4, 2004.
7. Ritala, M. Preface and the University of Helsinki papers in: *American Vacuum Society: ALD-04*. Helsinki, Finland: University of Helsinki, 2004 (Abstract Book/CD).

8. Aleskovsky, V. B. *Chem. Technol. Solids* 47, no. 10 (1974): 2145 Translated from Zhurnal Prikladnoi Khimii.
9. Ritala, M., and M. Leskela. "Atomic Layer Deposition." In *Handbook of Thin Film Materials*, Vol. 1, edited by H. Nalwa, San Diego: Academic Press, chap. 2.
10. Usui, A., and H. Sunakawa. *J. Phys. (Paris) Colloq.* C5 (1987): 48.
11. Tischler, M., and S. M. Bedair. "ALE of III–V Compounds." In *Handbook of Crystal Growth 3*, edited by D. T. J. Huerle, Amsterdam: Elsevier, 1994, chap. 4.
12. Kim, H. "Atomic Layer Deposition of Metal and Nitride Thin Films: Current Research Efforts and Applications for Semiconductor Device Processing." *J. Vac. Sci. Technol.* B26-(2003): 2231.
13. Hutchenson, D. *VLSI Forecast of ALD Equipment Revenues*, May 2005.
14. Bedair, S. M. "Atomic Layer Epitaxy Deposition Processes." *J. Vac. Sci. Technol.* B 12, no. 1 (1994): 179.
15. Ott, A. W. et al. "Al_2O_3 Thin Film Growth on Si (100) Using Binary Reaction Sequence Chemistry." *Thin Solid Films* 292, no. 1–2 (1997): 135.
16. Widjaja, Y., and C. Musgrave. "Quantum Chemical Study of the Mechanism of Aluminum Oxide Atomic Layer Deposition." *Appl. Phys. Lett.* 80 (2002): 18.
17. Ferguson, J. D. et al. "Atomic Layer Deposition of Ultrathin and Conformal Al_2O_3 Films on BN Particles." *Thin Solid Films* 371 (2000): 95.
18. Aarik, J. et al. "Phase Transformation of Hafnium Dioxide Thin Films Grown by Atomic Layer Deposition at High Temperatures." *Appl. Surf. Sci.* 173, no. 1–2 (2001): 15.
19. Ritala, M. et al. "Zirconium Dioxide Thin Films Deposited by ALE Using Zirconium Tetrachloride as a Precursor." *Appl. Surf. Sci.* 76 (1994): 436.
20. Zang, H. et al. "High Permitivity Thin Film Nanolaminates." *J. Appl. Phys.* 87, no. 4 (1921): 2000.
21. Yun, S.-J. et al. "Large-Area Atomic Layer Deposition and Characterization of Al_2O_3 Film Grown Using $AlCl_3$ and H_2O." *J. Korean Phys. Soc.* 33, no. S2 (1998): S170.
22. Aarik, J. et al. "Morphology and Structure of TiO_2 Thin Films Grown by Atomic Layer Deposition." *J. Cryst. Growth* 148, no. 3 (1995): 268.
23. Sammelselg, V. et al. "TiO_2 Thin Films by Atomic Layer Deposition, a Case of Uneven Films at Low Temperatures." *Appl. Surf. Sci.* 134, no. 1–4 (1998): 78.
24. Kattelus, H. P., and M. A. Nicolet. *Diffusion Phenomena in Thin Film and Microelectronic Materials*. edited by D. Gupta and P. S. Ho. Noyes, Park Ridge, NJ, 1989.
25. Hiltunen, L. et al. "Nitrides of Titanium, Niobium, Tantalum, and Molybdenum Grown as Thin Films by Atomic Layer Epitaxy Method." *Thin Solid films* 166 (1988): 149.
26. Klaus, J. W. et al. "Atomically Controlled Deposition of Tungsten and Tungsten Nitride Using Sequential Surface Reactions." *Appl. Surf. Sci.* 162 (2000): 479.
27. Fabreguettee, F. H. et al., "Demonstration of AlN ALD Using Hydrazine as the Nitrogen Precursor". San Jose, CA: American Vacuum Society, 2003 (ALD-03, Abstract Book/CD).
28. Juppo, M. et al. "Use of 1,1 Dimehylhydrizine in the Atomic Layer Deposition of Transition Metal Nitride Thin Films." *J. Electrochem. Soc.* 141 (2001): 3377.
29. Alen, P. et al. "*Tert*-Butyamine and Allylamine as Reductive Nitrogen Sources in Atomic Layer Deposition of TaN." *J. Mater. Res.* 17, no. 1 (2002): 107.
30. George, S. et al. "Atomic Layer Controlled Deposition of SiO_2 and Al_2O_3 Using ABAB…Binary Reaction Sequence Chemistry." *Appl. Surf. Sci.* 82/83 (1994): 460.
31. Klaus, J. W. et al. "Atomic-Layer Deposition of SiO_2 Using Catalyized and Uncatalyized Self-Limiting Surface Reactions." *Surf. Review and Letters* 6, no. 3–4 (1999): 435.
32. Yokohama, S. et al. "Atomic Layer Selective Deposition of Silicon Nitride on Hydrogen Terminated Surfaces." *Appl. Surf. Sci.* 130–132 (1998): 352.
33. Kim, Y. K. et al. "Novel Capacitor Technology for High Density Stand-Alone and Embedded DRAMs." *IEDM Tech. Dig.* (2000): 369 (IEEE Cat No: 00CH38138).
34. Ruhela, D. et al. "Low Temperature Deposition of AlN Films by an Alternate Supply of Trimethylaluminum and Ammonia." *Chem. Vap. Deposition* 2, no. 6 (1996): 277.

35. Haussman, D. et al. "Atomic Layer Deposition of Hafnium and Zirconium Oxides Using Metal Amide Precursors." *Chem. Mater.* 14 (2002): 4350.
36. Liu, X. et al. *ALD of HfO_2 Films from Tetrakis(Ethylmethylamino) Hafnium with OZONE and Water.* Seoul, Korea: American Vacuum Society, 2002 (ALD-02, Abstract Book/CD).
37. Dalton, et al., High performance ALD Reactor for Advanced Applications. Seoul, Korea: American Vacuum Society, 2006 (ALD-06, Abstract Book/ CD).
38. Kukli, K. et al., "Atomic Layer Deposition and Chemical Layer Deposition of Tantalum Oxide by Successive and Simultaneous Pulsing of Tantalum Ethoxide and Tantalum Chloride." *Chem. Mater.* 12 (2000): 1914: Haussman, D. et al, "Highly Conformal Atomic Layer Deposition of Tantalum Oxide Using Alkylamide Precursors." *Thin Solid Films* 443 (2003): 1.
39. Lee, J.-H. et al. "Mass Production Worthy HfO_2–Al_2O_3 Laminate Capacitor Technology Using Hf Liquid Precursor for Sub 100 nm DRAM." *IEDM Tech. Dig.* (2002): 221 (IEEE Cat No: 02CH37358).
40. Seidel, T. et al. *Progress and Challenges in ALD Applications*. San Jose, CA: American Vacuum Society ALD-03, 2003 (Abstract Book/CD).
41. Elliot, S. *Mechanisms of Ozone ALD.* 45. San Jose, CA: American Vacuum Society, 2005, ALD-05.
42. Liu, X. et al. "ALD of HfO_2 Films from Tetrakis(Ethymethylamino) Hafnium with Ozone." *J. Electrochem. Soc.* 152, no. 3 (2005): G213.
43. Karim, Z. et al. *Formation and Characterization of HfSiON for Advanced Semiconductor Devices*. San Jose, CA: American Vacuum Society, 2005 (ALD-05, Abstract Book page 87/CD).
44. Swerts, J. et al. *Engineering ALCVD TM HfSiO Gate Stacks for LSTP Applications*. San Jose, CA: American Vacuum Society, 2005 (ALD-05, Abstract Book page 63/CD).
45. Gordon, R. *Precursors with Metal–Nitrogen Bonds of ALD of Metals, Nitrides and Oxides*. San Jose, CA: American Vacuum Society, 2005 (ALD-05, Abstract Book page 10/CD).
46. Yokoyama, S. et al. "Self-Limiting Atomic-Layer Deposition of Si on SiO_2 by Alternate Supply of Si_2H_6 and $SiCl_4$." *Appl. Phys. Lett.* 79, no. 5 (2001): 617.
47. Klaus, J. W. et al. "Atomic Layer Deposition of Tungsten Using Sequential Surface Chemistry with a Sacrificial Stripping Reaction." *Thin Solid Films* 360 (2000): 145.
48. Aaltonen, T. et al. *Chem. Vap. Deposition* 9 (2003): 45; Aaltonen, T. et al. *ALD of Noble Metals-Exploration of the Low Limit of the Deposition Temperature*. Helsinki, Finland: American Vacuum Society, 2004, (ALD-04 Abstract Book/CD).
49. Nishizawa, J. et al. "Molecular Layer Epitaxy of Silicon." *J. Cryst. Growth* 99 (1990): 502.
50. Imai, S. et al. "Atomic Layer Epitaxy of Si Using Atomic H." *Thin Solid Films* 225 (1993): 168.
51. Rossnagel, S. et al. "Plasma Enhanced Atomic Layer Deposition of Ta and Ti for Interconnect Diffusion Barriers." *J. Vac. Sci. Technol.* B18 (2000): 2016.
52. Londergan, A. R. et al. "Process Optimization in Atomic Layer Deposition of High *K* Oxides for Advanced Gate Engineering". *Rapid Thermal and Other Short Time Processes III, Proc.* Vol. 2002–11. (The Electrochemical Society, Inc.).
53. Park, J.-S. et al. "Plasma-Enhanced Atomic Layer Deposition of Ta–N Thin Films." *J. Electrochem. Soc.* 149 (2002): C28.
54. Sneh, O. et al. "Atomic Layer Growth of SiO_2 on Si (100) Using $SiCl_4$ and H_2O in Binary Reaction Sequence." *Surf. Sci.* 334 (1995): 135.
55. Klaus, J. W. et al. "Growth of SiO_2 at Room Temperature with the Use of Catalyzed Sequential Half Reactions." *Science* 278 (1997): 5345; Klaus, J. W. "Growth of SiO_2 at Room Temperature with the Use of Catalyzed Sequential Half Reactions." *Surf. Sci.* 447 (2000): 81.
56. Haussman, D. et al. "Rapid Vapor Deposition of Highly Conformal Silica Nanolaminates." Science (2002): 298.
57. Lim, H.-S. et al. "Analysis of a Transient Region during the Initial Stage of Atomic Layer Deposition." *J. Appl. Phys. Part 1* (2000).
58. Elam, J. W. et al. "Surface Chemistry and Film Growth during TiN Atomic Layer Deposition Using TDMAT and NH_3." *Thin Solid Films* 436 (2003): 145.

59. Matero, R. et al. "Effect of Water Dose on the Atomic Layer Deposition Rate of Oxide Thin Films." *Thin Solid Films* 386 (2000): 1.
60. Karim, Z. et al. "Advanced Metal Gate Electrode Options Compatible with ALD and AVD HfSiOx-based Gate Dielectric", *ECS Transactions*, 3, no. 3 (2006): 363.
61. George, S. "AVS Short Course on Atomic Layer Deposition (ALD)." *Lecture Notes* (2004).
62. Puurunen, R. L. *Atomic-Scale Modeling of Atomic Layer Deposition Processes*. San Jose, CA: American Vacuum Society, 2005 (ALD-05, Abstract Book page 16/CD).
63. Green, M. et al. *J. Appl. Phys.* 92 (2002): 7168 (see also, A New Model of Atomic Layer Deposition, and Its Applications to the Nucleation and Growth of HfO_2 Gate Dielectric Layers. American Vacuum Society ALD-03. Abstract Book/CD, San Jose, CA 2003).
64. Gusev, E. P. et al. "Ultrathin HfO_2 Films Grown on Silicon by Atomic Layer Deposition for Advanced Gate Dielectrics Applications." *Microelectron. Eng.* 69 (2003): 145.
65. Bresling, W. F. A. et al. "Characterization of ALCVD- Al_2O_3–ZrO_2 Nanolaminates, Link between Electrical and Structural Properties." *J. Non-Cryst. Solids* 303 (2002): 123.
66. Frank, M. M. et al. "Nucleation and Interface Formation Mechanisms in Atomic Layer Deposition of Gate Oxides." *Appl. Phys. Lett.* 82 (2003): 4758.
67. Elam, J. W. et al. "Nucleation and Growth During Tungsten Atomic Layer Deposition on Oxide Surfaces." *Thin Solid films* 386 (2001): 41.
68. Elam, J. W. et al. "Improved Nucleation on TiN ALD Films on SILK Low-*k* Polymer Dielectric Using an Al_2O_3 ALD Adhesion Layer." *J. Vac. Sci. Technol.* B21 (2003): 1099.
69. Sneh, O. et al. "Thin Film Atomic Layer Deposition Equipment for Semiconductor Processing." *Thin Solid Films* 402 (2002): 248.
70. A.R. Londergan, et al, "Pathways for Advanced Transistor Using Hafnium-Based Oxides by Atomic Layer Deposition." In *Advanced Short-Time Thermal Processing for Si-Based CMOS Devices, Proceeding of the International Symposium*, Proc. Vol. 2003–14. (The Electrochemical Society, Inc.)
71. Lee, J. H. et al. "Practical Next Generation Solution for Stand-Alone and Embedded DRAM Capacitor." *VLSI Tech. Dig.* (2002) (IEEE Cat No: 02CH37302).
72. Tan, Y. N. et al. "High *K* HfAlO Charge Trapping in SONOS Type Nonvolatile Memory for High Speed Operation." *IEDM Tech. Dig.* (2004): 889 (IEEE Cat No: 04CH37602).
73. Seidel, T. et al. "Characterization of Charge Modification in ALD Nanolaminates." *ECS Symposium* October 17, (2005).
74. Kim, S.-J. et al. "Engineering of Voltage Nonlinearity in High-*K* MIM Capacitor for Analog/Mixed-Signal IC's." *IEDM Tech. Dig.* (2004).
75. Zaitsu, S. et al. "Optical Thin Films Consisting of Nanoscale Laminated Layers." *Appl. Phys. Lett.* 80 (2002): 2442.
76. Gusev, E. P. et al. "Ultrathin Hi k Dielectrics Grown by Atomic Layer Deposition: A Comparative Study of ZrO_2, HfO_2, Y_2O_3 and Al_2O_3." *Proc. ECS* 97 (2001): 189.
77. Zhao, C. et al. "Thermo Stability of Amorphous Zirconium Aluminate High-*k* Layers." *J. Non-Cryst. Solids* 303 (2002): 144.
78. Kukli, K. et al. "Properties of Ta_2O_5-Based Dielectric Nanolamoinates Deposited by Atomic Layer Epitaxy." *J. Electrochem. Soc.* 144 (1999): 300.
79. Kim, Y.-S. et al. "Multilayered Tantalum–Aluminum Oxide Films Grown by Atomic Layer Deposition." *J. Korean Phys. Soc.* 35, no. S2 (1999): S216.
80. Kumagai, H. et al. "Titanium Oxide/Aluminumoxide Multilayer Reflectors for Water Window Wavelength." *Appl. Phys. Lett.* 70, no. 18 (1997): 2338.
81. Kanniainen, T. et al. "Growth of Dielectric HfO_2/Ta_2O_5 Thin Film Nanolaminate Capacitors by Atomic Layer Epitaxy." *Proc. ECS* 97–31 (1998): 36.
82. Londergan, A. R. et al., *Engineered Low Resistivity Titanium–Tantalum Nitride Films by Atomic Layer Deposition*. San Francisco, CA: MRS Spring Meeting, 2001.
83. Cho, B. C. et al., *Atomic Layer Deposition of TiAlN for High K Applications*. San Jose, CA: American Vacuum Society, 2005 (ALD-05, Abstract Book page 84/CD).
84. Kim, G. Y. et al. "A High Deposition Rate Process Using Limited Optimized Reaction ALD." Helsinki, Finland, 2004 (ALD-04 Abstract Book/CD).

85. Many ALD system descriptions may be found on the United States Patent and Trademark Office (USPTO) website. Interested readers can search, for example, using "apparatus" and "ALD".
86. Glime, W. *Design, Testing and Manufacture of Fast Switching Valves for High Productivity Modes of ALD*. San Jose, CA: American Vacuum Society, 2005 (ALD-05, Abstract Book page 102/CD).
87. Shatas, S., and Ono, Y. Steve George-private communications.
88. Sneh, O. *High Productivity ALD Using Synchronously Modulated Flow and Draw*. Helsinki, Finland: American Vacuum Society, 2004 (ALD-04 Abstract Book/CD).
89. Gordon, R. et al. "A Kinetic Model for Step Coverage by Atomic Layer Deposition in Narrow Holes or Trenches." *Chem. Vap. Deposition* 9, no. 2 (2003): 73.
90. Kim, K. "Technology for sub-50 nm DRAM and NAND Flash Manufacturing." *IEDM Tech. Dig. 13.5.1* (2005): 333 (IEEE Cat No: 05CH37703).
91. Gutsche, M. et al. "Capacitance Enhancement Techniques for sub-100 nm Trench DRAMs." *IEDM Tech. Dig.* (2001): 411 (IEEE Cat No: 02CH37303).
92. Lutzen, J. et al. "Integration of Capacitor for Sub-100-nm DRAM Trench Technology." *VLSI Tech. Dig.* (2002): 178 (IEEE Cat No: 02CH37302).
93. Koyama, M. et al. "Effects of Nitrogen in HfSiON Gate Dielectric on the Electrical and Thermal Characteristics." *IEDM Tech. Dig.* (2002): 849 (IEEE Cat No: 02CH37358).
94. Rotondaro, A. L. P. et al. "Advanced CMOS Transistors with a Novel HfSiON Gate Dielectric." *VLSI Tech. Dig.* (2002): 148 (IEEE Cat No: 02CH37302).
95. Jung, H. S. et al. "Improved Current Performance of CMOSFETs with Nitrogen Incorporated HfO_2–Al_2O_3 Laminate Gate Dielectric." *IEDM Tech. Dig.* (2002): 853 (IEEE Cat No: 02CH37358).
96. Lee, J. H. et al. "Poly-Si Gate CMOSFETs with HfO_2–Al_2O_3 Laminate Gate for Low Power Applications." *VLSI Tech. Dig.* (2002): 84 (IEEE Cat No: 02CH37302).
97. Borland, J. O. et al. *Meeting Challenges For Engineering the Gate Stack Solid StateTechnology*. July, 2005.
98. Ivanova, A. R. et al. "Integration of PECVD Tungsten Nitride as a Barrier Layer for Copper Metallization." *Mat. Res. Soc. Symp. Proc.* 564 (1999): 321.
99. Rossnagel, S. et al. *The Impact of ALD in Semiconductor Interconnects and Contacts*. Helsinki, Finland: American Vacuum Society, 2004 (ALD-04, Abstract Book/CD).
100. Cheong, S. H. et al. *The Evaluation of ALD-WN/W Process for Sub-70 nm Contact Plug Technology*. San Jose, CA: American Vacuum Society, 2005 (ALD-05, Abstract Book page 85/CD).
101. Lu, Q. et al. "Molybdenum Metal Gate MOS Technology for Post-SiO2 Gate Dielectrics." *IEDM Tech. Dig.* (2000): 641 (IEEE Cat No: 00CH38138).

15

Physical Vapor Deposition

15.1 Introduction and Semiconductor Applications **15**-1
15.2 Sputtering Background and Basics **15**-2
15.3 PVD Systems **15**-6
15.4 Applications and Variations for Interconnect Applications **15**-9
Planar, Conductive, or ARC Films • Reflow and Surface Mobility-Based Deposition • Directional Deposition • Ionized Deposition
15.5 Summary, Future Directions **15**-24
References **15**-25

Stephen M. Rossnagel
IBM Thomas J. Watson Research Center

15.1 Introduction and Semiconductor Applications

Physical vapor deposition (PVD) thin film technology covers a rather broad range of deposition techniques. The general feature that describes PVD is that films are deposited atomically by means of fluxes of individual neutral or ionic species. This differentiates them from chemical vapor deposition (CVD), in which films are precipitated from the gas phase by a chemical reaction, and also from electrodeposition, in which atoms or ions in an aqueous solution are plated onto a surface.

PVD techniques include all techniques based on evaporative deposition, such as e-beam or hot-boat evaporation, reactive evaporation and ion plating. PVD techniques also include all processes based on sputtering, either by a plasma or by an ion beam of some sort. PVD is also used to describe deposition from arc sources which may or may not be filtered.

In the current day semiconductor industry, PVD technology is entirely based on physical sputtering, usually using a specific type of diode source known as a magnetron. Atoms are physically sputtered from the magnetron cathode by means of a local plasma, and the sputtered metal atoms are then used as the basis for film deposition once they travel to the location of the sample. The other PVD techniques, such as e-beam evaporation or ion beam sputtering, have been used in either the earlier days of semiconductor processing, or else exist on the fringe of development projects, and do not have any significant relevance to mainstream interconnect processing.

The primary semiconductor applications for PVD technology are the deposition of metal and compound lines, pads, vias, contacts, and related connections which are used to connect with the junctions and devices present on the Si wafer surface. The emphasis here is on electrical connection: the PVD features are not part of the semiconductor junction directly, but make the electrical connections between it and nearby circuit features.

With the feature size of the desired metal structures decreasing with each succeeding semiconductor generation, the techniques to both pattern and deposit PVD features has evolved over the years. Aside

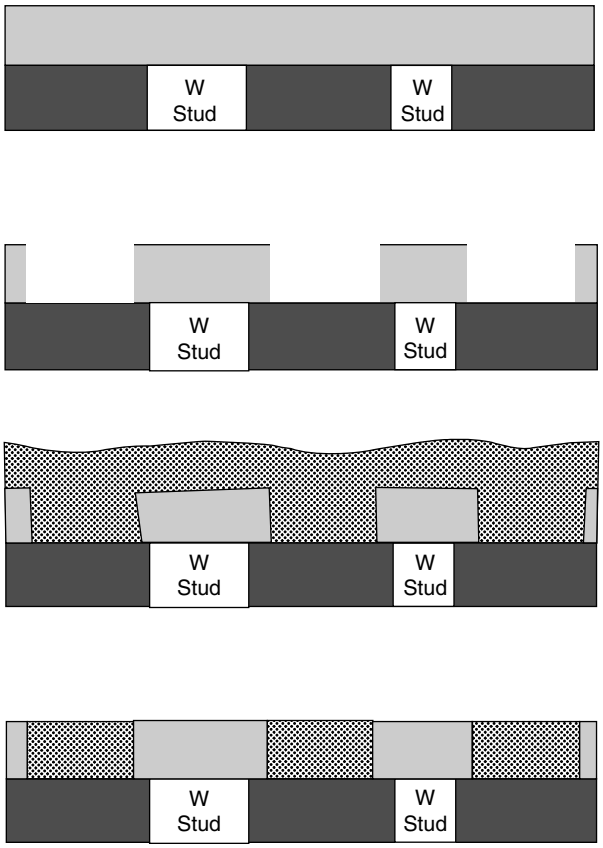


FIGURE 15.1 Sketch of several steps of the reactive ion etch-metal patterning process.

from the very early days of photoresistbased lift-off deposition, there have been two primary lithographic pattern technologies used for PVD deposition. The first is a subtractive process based on the deposition of planar films and the subsequent patterning of features by means of reactive ion etch (RIE) [1]. This will be called the "RIE-metal" technology for this chapter (Figure 15.1).

The second general class of deposition techniques uses an alternate approach where trenches and vias are etched into planar, dielectric film layers. These features are then filled with metal and polished flat, in a technique known in the industry as "Damascene" processing, named after similar ancient jewelry making techniques (Figure 15.2) [2].

The PVD technologies used for each technology: RIE-metal, and Damascene, are completely different, as in one case there is a desire for a completely planar, flat film covering over small steps and bumps, and in the second case there is a need for a more directional or preferential deposition where the deposition occurs deep into features such as trenches and vias.

15.2 Sputtering Background and Basics

For PVD techniques based on sputtering, the vast majority of the cases of interest will use bombardment of a negatively-biased cathode by means of a high energy, inert gas ion, typically Ar, but occasionally other inert gas species (Ne, Kr) or also occasionally reactive species such as oxygen or nitrogen. The physical sputtering process is well understood in the literature [3–8] and consists of a sequence of

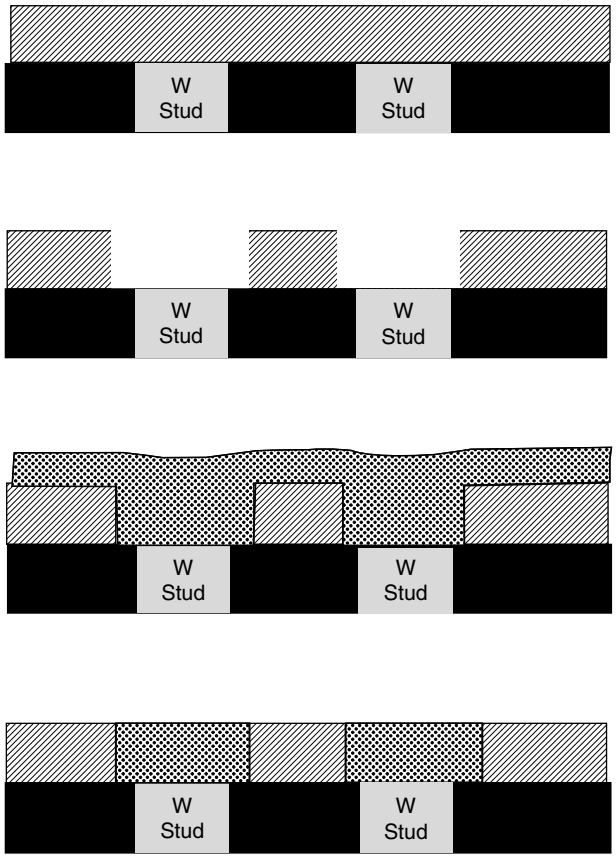


FIGURE 15.2 Sketch of steps used for Damascene metal filling technology.

energetic, violent collisions between the incident particle and a cold lattice of the target (Figure 15.3). The effect of this incoming ion is to physically dislodge one or more of the target atoms, which then in turn move on striking other atoms within the surface lattice structure. This near-cascade of collisions can eventually result in one or more atoms in the near-surface layers having just enough kinetic energy (and the appropriate direction) to overcome the surface binding energy and escape. This escaping atom is then described as having being sputtered from the surface, although it may have originated 1–2 layers down from the original surface.

The exact sequence of collisions is, as might be evident from the sketch, very dependent on the exact trajectory and impact site of the incident ion. Since these are not controllable features, sputtering is usually described in terms of average effects: the result of many millions, or more impacts by energetic ions, and the average emission of sputtered particles from variety of lattice orientations. This is known generically as the sputter yield, and it is simply the ratio between the number of emitted, sputtered particles and the number of incident, high energy impacting ions. The sputter yield is generally a number which is based on both the kinetic energy of the incoming ion (as well as it mass) along with the species of the impacted surface. The yield varies from essentially undetectable at very low ion energies (10 s of eV) to numbers on the order of 1–5 at modest ion energies of many hundred to thousands of electron volt. A graph of the sputter yields for some materials of interest to semiconductor processing is shown in Figure 15.4.

The angular emission distribution of the sputtered atoms is characterized by a so-called cosine distribution in which the relative flux at any angle other than the surface normal scales with the

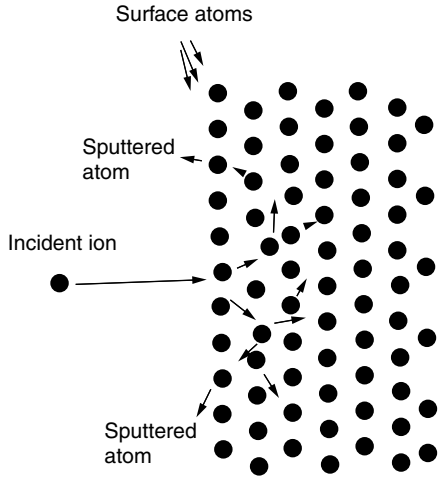


FIGURE 15.3 Schematic of sputtering event. (From Rossnagel, S. M., In *Handbook of Vacuum Science and Technology*, Academic Press, Orlando, FL, (1997): 609.)

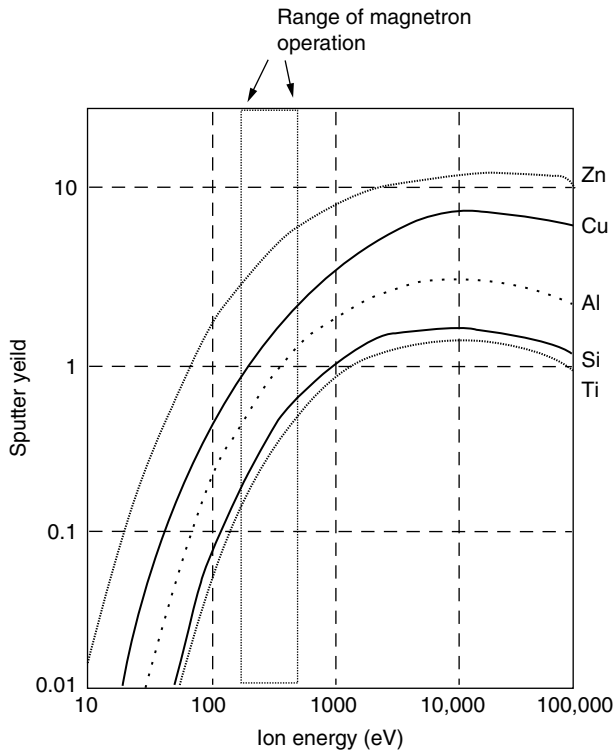


FIGURE 15.4 Sputter yields as a function of incident. Ar ion energy for several materials. (From Rossnagel, S. M., In *Handbook of Vacuum Science and Technology*, Academic Press, Orlando, FL, (1997): 609.)

normal-direction flux times the cosine of the angle from normal. There are numerous deviations from this general rule, and in general, lower ion energies tend to result in a flatter, "under-cosine" distribution, whereas higher ion energies tend to result in a more peaked, over-cosine distribution [9]. There are also unusual cases where the crystal structure of the sputtered material can have an effect on the angular emission distribution. This was first observed in the 1950s by Wehner, who saw spots in the emission distribution characteristic of the underlying crystal structure [10]. More recently, preferential emission has been incorporated into magnetron target fabrication to produce and emission distribution, which is much more peaked than the conventional cosine distribution [11].

The sputtered atoms must then move through the background gas to land at the desired sample surface. Most magnetron sputter deposition systems are operated at pressures in the low milliTorr range, where the mean free path for gas-phase collisions is generally greater than the distance between the cathode and the sample, or throw distance. This means that most of the sputtered atoms will have ballistic trajectories with few, if any, in-flight collisions. At pressures above a few milliTorr, this no longer holds, and as the pressure approaches 30 mTorr or so, essentially all of the sputtered atoms have numerous gas-phase collisions and lose essentially all of their original kinetic energy and direction from the sputtering process [12–16]. This process is known as 'thermalization' from the point of view of the sputtered atom, which becomes thermally equilibrated with the background gas. The process also results in significant as heating, though, resulting in a local rarefaction of the background gas in the region in from of the cathode. For significant levels of applied magnetron power, the resulting gas density can be as low as 20% of the starting density, with an equivalent gas temperature of 1500 K or more [17]. This sputter-induced change to the gas density can also be used in a model of the plasma discharge to predict the current–voltage trends of the cathode [18].

The efficiency of the transport process is partly dominated by operating conditions and partly by the design and configuration of the system. Since the emission process is essentially a cosine distribution and in most production tools the cathode diameter is not much smaller than the internal chamber diameter, significant deposition will occur on the chamber sidewalls as well as any fixtures such as shuttered, ground shields, clamp rings, etc. The probability of transport can be characterized by a number between 0 and 1, where 1 means that all sputtered atoms from the cathode land on the desired sample surface and 0 implies that no atoms are deposited [17]. Although rarely measured, the data shows expected trends with pressure, throw distance and gas used, as well as target species (Table 15.1). In this latter case, it is expected that when the target atomic weight exceeds the weight of the background gas, transport will be more efficient.

A more commonly used metric to characterize deposition efficiency is to calculate the unit deposition rate per watt of applied power. This is quite sensitive to the system used. The results are usually given in units of Angstroms/sec/ Watt. An example of this type of data is shown in Table 15.2 for an Applied

TABLE 15.1 Deposition Probability for Magnetron Sputtering at 1000 W, 200 mm dia Cathode Planar Magnetron

Throw (cm)	Pressure (mtorr)	Deposition Probability
Ar Sputtering of Cu		
5	5, 20, 30	0.63, 0.49, 0.54
9.5	5, 20, 30	0.48, 0.47, 0.45
14.5	5, 20, 30	0.39, 0.35, 0.31
Ne Sputtering of Al		
5	5, 20, 30	0.80, 0.56, 0.52
9.5	5, 20, 30	0.40, 0.42, 0.40
Ar Sputtering of Al		
5	5, 20, 30	0.60, 0.46, 0.42
9.5	5, 20, 30	0.44, 0.45, 0.35
Kr Sputtering of Al		
5	5, 20, 30	0.52, 0.45, 0.38
9.5	5, 20, 30	0.35, 0.27, 0.22

Source: From Rossnagel, S. M., *J. Vac. Sci. Technol.*, A6, (1988): 19.

TABLE 15.2 Deposition Rates per Watt and also Deposition Probability for Applied Materials Endura Sputtering System

Materials	Power (kW)	Rate (Å/min/W)
AlCu(0.5)	12.7	1 (new cathode) 0.75 (old cathode)
Ti	1	0.17
TiN (nitride mode)	4	0.15
Ti (collimated, 1.5:1)	7	0.043

The cathode is 12.98 in dia (Standard 200 mm size), the throw distance is 5 cm, and the samples are 200 mm Si wafers.

Source: From Rossnagel, S. M., *J. Vac. Sci. Technol.*, B16, (1998): 2585.

Materials Endura sputtering chamber (12.98 in cathode, 5 cm throw) [8,19]. In general, a typical number for this deposition efficiency is on the order of 1 Angstrom/sec/Watt, with higher numbers for materials such as Cu which have a high sputter yield.

15.3 PVD Systems

PVD technology covers a lot of various non-chemical deposition techniques. For semiconductor applications, essentially 100% of the deposition systems are based on a variation of a dc-diode device known as a magnetron. There are occasionally reports of work done using rf-diode deposition systems, but these are usually used in cases of dielectric materials, such as the high-k dielectric, and are not in widespread usage. Readers interested in rf technology are referred to a review chapter by Logan of IBM [21].

A magnetron cathode differs from a conventional, planer cathode in that there is a local magnetic field parallel to the cathode surface. This is shown in Figure 15.5. The effect of the tangential *B* field is such that secondary electrons which are emitted from the cathode surface due to ion bombardment (which is what

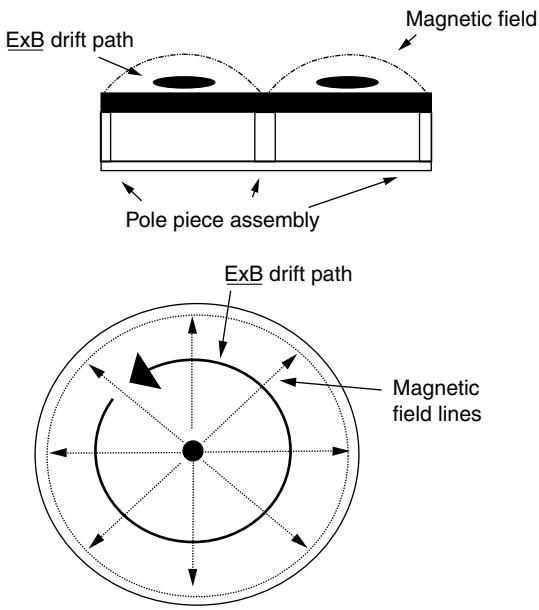


FIGURE 15.5 Schematic of a circular, planar magnetron cathode.

causes the sputtering), undergo an "$E \times B$" drift around the cathode surface in a manner similar to the Hall Effect. These drifting electrons are trapped close to the cathode region and can lead to very high levels of gas ionization, which results in very large discharge (ion) currents.

The magnetrons used in semiconductor production systems derive from this basic design. The cathode diameter is typically 50% larger than the sample to be deposited on (30–32 cm for a 200 mm wafer sample). This scaling is likely to hold as wafers migrate to the 300 mm generation, resulting in cathode diameters on the order of 45 cm. The cathode-to-sample distance, or 'throw' distance ranges from about 3 to 10 cm, with most tools operated at about 5 cm. In most production tools, the waters are positioned horizontally, such that they are facing up, and the magnetrons are configured to sputter down onto the wafer. Originally there was concern that this configuration would result in the maximum particulate contamination of the sample, due simply to gravity. However, most sub-micron particles in vacuum systems are much more influenced by static charges, van-der Wahl forces, and gas-phase turbulence than they are by gravity.

The wafers are fixed on the substrate platforms in one of three ways: a physical clamp ring held down by either springs or gravity, no clamping at all with the wafer sitting up on small pins or pads, and electrostatic clamping. The use of clamp rings reduces the useful area of the wafer by several percent but tends to physically and thermally couple the wafer much better to the pedestal, giving moderate temperature control. Clampless operation gives full wafer coverage, but at the expense of any control whatsoever over the thermal or electrical condition of the wafer. Electrostatic clamping can give the best of both worlds, and many major tool suppliers are beginning to make low temperature, biasable e-chucks available [22]. A continuing concern over the use of electrostatic chucks is to be able to diagnose the no-wafer failure mechanism. This means a case where for some reason the wafer does not arrive in the deposition chamber and yet the deposition is turned on. Due to the electrical nature of the electrostatic chuck, a blanket metal deposition onto an uncovered chuck results in destruction of the chuck. A second e-chuck concern is the presence of backside particulates, which may be pushed into the backside by the strong chucking force. This can lead to significant chamber-to-chamber contamination.

Production magnetrons are configured with a moving magnet array, rather than the fixed magnets of Figure 15.5. As shown in Figure 15.6, the magnet assemble is located behind the cathode surface and sloshes around in the water cooling bath, rotating about the cathode centerline. The etch track is usually somewhat heart-shaped with the indentation at the top of the heart roughly on the cathode centerline. This etch track/magnet system is driven by an external motor to cycle around the cathode surface at several Hertz. The exact shape of the etch track can be tailored by adjusting the magnets or the pole pieces. Changing the shape of the etch track results in changes in the uniformity of the deposition, which may be desirable either because of the material used or else the specific geometrical configuration. For example, if the throw distance is increased then more sputtered atoms are lost to the sides of the chamber, resulting in a net reduction in rate as a function of radius. In this case, it may be necessary to adjust the etch track to increase the erosion rate near the edge of the cathode to compensate. Alternatively, the use of a collimator (described below) completely changes the requires erosion profile of the cathode. Each cell of the collimator functions as a tiny pinhole camera, imaging a surface of the cathode onto the sample. Therefore the etch uniformity in that case must be higher than the long-throw case. In each case, it is also possible to tailor the erosion uniformity of the cathode to provide a high level of cathode utilization. Rotating-magnet cathodes might typically use 50–70% of the high purity cathode material before they must be replaced. This is much different from the conventional cathode (Figure 15.5), which may only 15% of its material before the erosion track becomes too deep. Higher cathode utilization results in lower operating cost for the tool as well as longer times between cathode changes.

All magnetron cathodes are water cooled and the amount of cooling turns out to be the practical limit to discharge operation. This is quite unlike most other types of plasma devices, but in a magnetron the current–voltage interaction is such that additional discharge power is easy to add to the cathode in the form of amperes of ion current. The practical limit is when the cathode deforms or even melts. Practical and safety concerns limit the temperature of the cooling water at the cathode to 60°C or so. Assuming excellent heat transfer, this results in a requirement for about 1/4 gallon-per-minute of water flow for

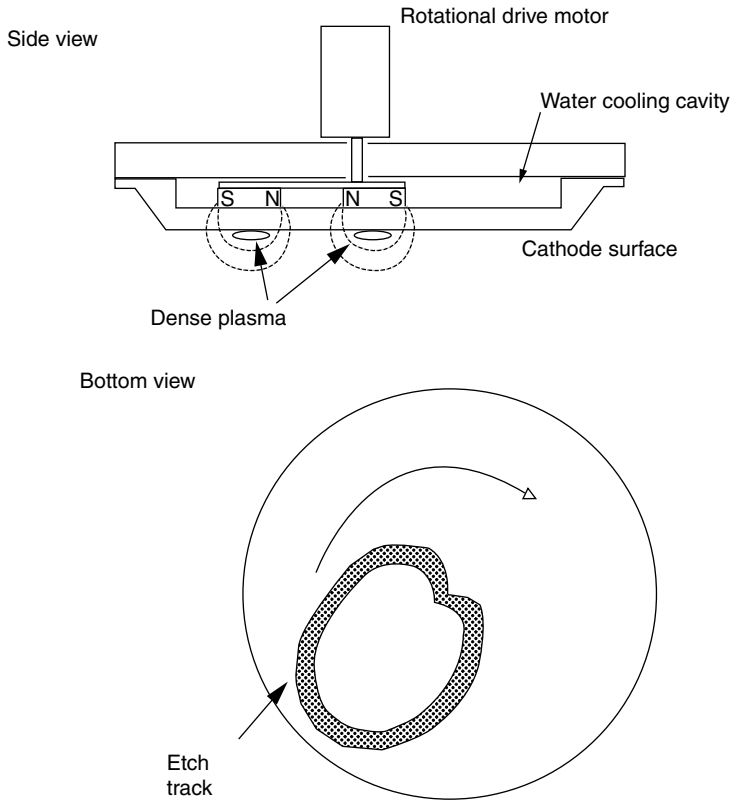


FIGURE 15.6 Magnetron design with moving-magnet, heart-shaped $E \times B$ etch track. (From Powell, R. A. and S. M. Rossnagel, *PVD for Microelectronics*, Academic Press, Boston, NY, 1998.)

each kilowatt of applied discharge power. It has often been said that a magnetron sputter deposition system is simply a very expensive water heater.

Most commercial magnetron systems use a removable target-and-backing-plate assembly, which is bolted to a non-conducting housing on which the motor drive is mounted. The cavity behind the cathode contains the magnets and also the cooling water, typically with water line of about 2 cm diameter. Recent cathode designs have migrated towards a target configuration, which has built-in water channels (Figure 15.7) [23]. This scales better to larger cathode size and allows the magnets to be operated in air rather than immersed in water, reducing the inevitable corrosion of both the magnets and the pole pieces. It does, however, make the target assemble more expensive. A conventional target-and-backing plate assembly for a 200-mm sample system might cost $3000 or so for moderate-purity AlCu or Ti.

The composition and purity of the target are obviously quite important to the deposition application. In the semiconductor industry, target composition is almost always described in the form of weight percent, which can differ significantly from atomic percent, which is used in traditional chemical analysis or chemical formulas. For example, $Ti_{0.8}$ $W_{0.2}$ is a compound composed of 80% Ti atoms and 20% W atoms. However, since the atomic weight of W is 3.8× that of Ti, the weight percent of this compound is Ti(50)W(50), where the general nomenclature is that weight percents are written with parenthesis. To further complicate the description, if one component in an alloy or compound is allowed with a trace amount of another material, the dominant percentage is often dropped. An example of this is AlCu(0.5), which has 0.5% by weight of Cu and 99.5% by weight of Al.

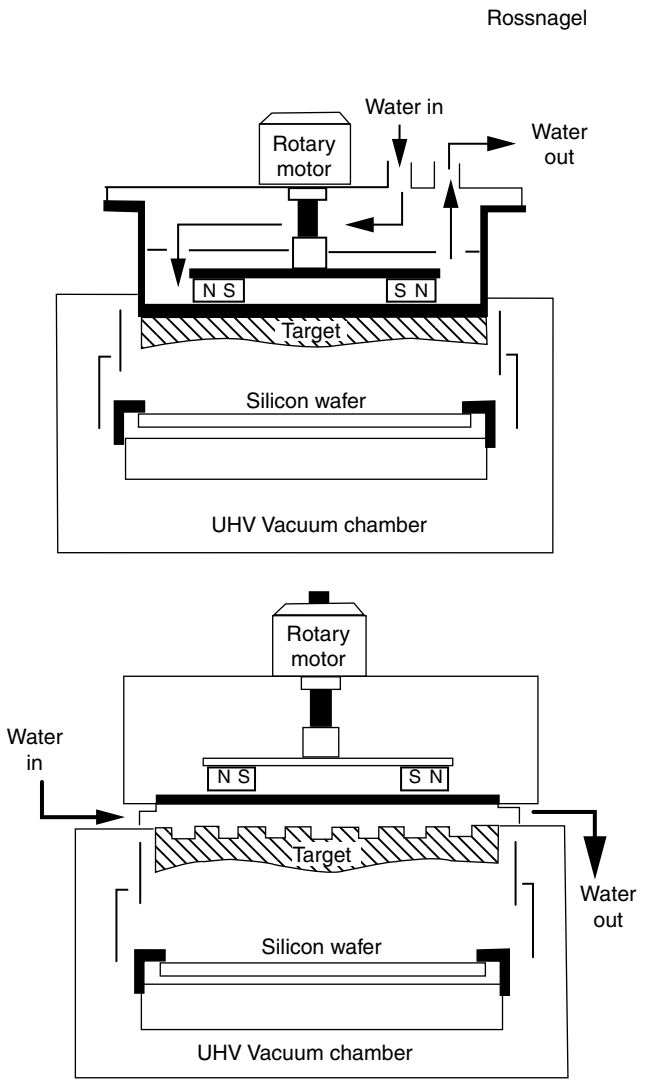


FIGURE 15.7 Magnetron cathode design with integrated water cooling channels. (From Powell, R. A. and S. M. Rossnagel, *PVD for Microelectronics*, Academic Press, Boston, NY, 1998.)

Target purity is described in terms of percent-purity, and from a practical point of view is always less than 100%. Typical purities for Al are 99.999% (5 9's) to 99.9995% (5 9's five), with lower purities for Ti(4–5 9's) and higher purities for Cu(6–7 9's). The purity for refractory materials is generally lower (3–4 9's), and these materials are not in widespread usage in semiconductor processing, yet. In all cases, though, a high purity disk of the desired material is typically diffusion bonded to an Al or Cu backing plate, which is bolted to the magnetron cathode.

15.4 Applications and Variations for Interconnect Applications

Semiconductor applications of PVD films tend to take two general approaches: planar films which are then etched using RIE technology, and fill-like applications, where the sample has trenches and/or vias etched into a planar surface.

15.4.1 Planar, Conductive, or ARC Films

Perhaps the widest usage of PVD films in the semiconductor industry is simply the blanket deposition of AlCu(0.5) metal layers. These films are deposited over planar surfaces or else surfaces with very low aspect ratio features. It is critical that the film be as flat as possible because the subsequent step is photolithographic exposure with a very shallow depth of field in preparation for subtractive etching. Typically, AlCu films are deposited at modest rate (a few thousand angstroms per minute) with a low level of wafer heating. This leads to grain growth in the films such that the grain size is comparable with the film thickness (1/2 μm level). The film orientation is typically (111).

A second widespread application of planar films is the deposition of TiN to form an anti-reflection coating (ARC). This tailored film is used to optimize the photolithographic process, and may also play a role as either an etch stop or protective layer. Typically TiN films are deposited from pure Ti targets sputtered with a partial pressure of nitrogen. The nitrogen reacts with the Ti film atoms (at the sample surface) if the sample temperature is adequate, typically 250°C or so.

15.4.2 Reflow and Surface Mobility-Based Deposition

The fundamental problem of putting atoms into a deep feature can be solved in atleast 2 ways: using either enhanced surface mobility of the deposited atoms, or else enhanced directionality of the atoms as they are deposited. The first mode is closest to the planar film applications discussed above and uses essentially the same toolset, although it does not address the fundamental, non-directional nature of sputter deposition. Two general techniques have evolved for addressing surface mobility of atoms: the first is simply increased sample temperature, and the second is based on a more macroscopic extrusion of film material into deep features.

The effects of increased sample temperature on film deposition include increased surface diffusion, grain formation and growth, increased chemical activity of the film material, and also somewhat negative features such as film agglomeration, void formation, re-evaporation, precipitation, and interdiffusion of the film with underlying layers. The goal of a thermal reflow process is that atoms move from the planar or field areas of the sample towards inset or deep features, such as trenches or vias. Movement of these atoms means that their original deposition trajectory (and conditions) are unimportant, and this is consistent with conventional, high rate PVD magnetron deposition technology. Since the bottom of a via or trench has a concave shape, this will tend to be a sink for thermally diffusing atoms. However, this assumes that the trench remains open to the top during the reflow process. If the top is closed and a void formed, then subsequent atom motion is by means of bulk diffusion, rather than surface diffusion, and is characterized by a much higher activation energy. This means that at any given sample temperature, a bulk-diffusion-dominated process will be significantly slower than a surface diffusion process.

Thermal reflow was first applied to semiconductor filling in the late 1980s [24], and showed significant filling of low aspect ratio features in Al at temperatures around 500°C. It is generally required for reflow technology that the first layers of the film wet the underlying surface and have good adhesion, or else the effect of additional sample temperature will form clusters and droplets, rather than a continuous film. This is generally done with a 2-step process, in which the seed layer is deposited at low temperature and the sample temperature is increased such that the remainder of the deposition is at high enough temperature to facilitate rapid reflow [23]. Several variations on this process include the use of collimated sputtering or long-throw deposition or even the use of a CVD layer for the seed layer. The latter process has been commercialized and may be advantageous in slightly reducing the high temperature required for reflow [25]. This physics behind any sort of advantage have yet to be explained, but may be related to subtle chemical effects, such as the presence of hydrogen in the film deposited by CVD.

Thermal reflow processes require high levels of cleanliness because surface contamination of the wafer or gas phase impurities such as oxygen or water may significantly impede the surface diffusion process. For the case of reflow Al, very small amounts of oxygen (pressure of 10^{-7} Torr) are sufficient to form small oxide islands with then impede diffusion. Thin layers of Ti deposited just prior to Al reflow can

result in better wetting and reflow of the Al as well as better adhesion [26]. The Ti can then be incorporated into the Al as $TiAl_3$ due to the elevated temperature, and this can serve to reduce film stress and the possibility of electromigration-induced failure [23]. This $TiAl_3$ phase, however, has a high resistivity which can lead to increased line resistance.

Reflow processes have been successfully applied to both the Al and Cu interconnect systems [27], although several intrinsic concerns have become apparent. The first concern relates to the requirement that the via or trench remain open during the reflow process, such that the activation energy for surface diffusion, as opposed to bulk diffusion, is dominant. This limits both the deposition rate as well as the minimum size of the feature. The interplay is between the non-normal incidence of the sputter deposition process, which tends to form voids, and the rate of surface mobility, which tends to fill the smallest vias and trenches first and also keep them open. The tradeoff is determined by the highest acceptable substrate temperature, as high temperature results is faster diffusion. Typically, the rule of thumb for oxide-based semiconductor processing is a maximum temperature of 400°C, and that temperature will decline significantly as the newer, low-*k* dielectrics are introduced.

The second concern relates to both the size of the feature as well as the surface density of features on the wafer. Since it simply takes more time, and more atoms, to fill a large feature as opposed to a small one, the larger features will lag during the processing. This is, perhaps, not a critical problem because filling low aspect ratio large features is not difficult using conventional PVD. In terms of feature density, though, the features in the middle of large patterns of features will tend to fill last, simply because the supply of atoms is from the side areas, and these will be captured first in the outermost features in an array.

A variation on reflow technology that has been developed recently is known generically as 'high pressure filling' [28]. This process uses conventional sputtering of blanket films, followed by exposure of those films to extremely high static gas pressures of an inert gas, such as Ar. The key to this process is, unlike conventional reflow deposition, it is greatly desired to deposit the films in such a way that voids are formed within high aspect ratio features. The samples are then removed from the sputtering chamber and introduced into a high pressure chamber. The temperature is raised to around 400°C, and Ar is introduced into the chamber at a level of 600–700 atmospheres. The large pressure, coupled with the elastic nature of the Al at 400°C (0.75 of melting T), allows the Al films to be squeezed or pushed down into the vias (Figure 15.8). There remains only a few milliTorr of inert gas within the void, and this gas is incorporated in the final structure at the 0.1 ppm level.

15.4.3 Directional Deposition

15.4.3.1 Long Throw

Most PVD systems are designed for maximum rate, and have short throw distances. This also results in the fewest number of atoms lost to the chamber walls. By moving the sample farther away from the cathode an increasing fraction of the sputtered atoms are lost onto the sidewalls of the chamber [29–31]. This results in a net reduction in the deposition rate at the sample, and also results in a net change in the average directionality of the depositing atoms. Atoms which are sputtered from the target surface at low angles (i.e., far from normal incidence) are more likely to land on the chamber sidewalls than on the wafer sample. The atoms arriving at the sample are more likely to be closer to normal incidence than the conventional, short-throw deposition. This geometrical filtering process is known generically as 'long throw' sputter deposition. The process is limited in a practical sense by the operating pressure of the system and gas scattering. To reduce in-flight scattering, the mean free path for the sputtered atoms should exceed the throw distance. To have any significant degree of directional filtering, the throw distance needs to be on the order of the cathode diameter, of 25 cm for a 200-mm wafer system. This places a practical pressure limit of a few tenths of a milliTorr on the operating pressure, as higher pressures will result in shorter mean free path distances than the throw distance.

Manufacturing applications of long throw deposition tend to have throw distances of about 25 cm, which limits the depositing flux to about $\pm 45°$. Greater directionality can only be obtained with longer

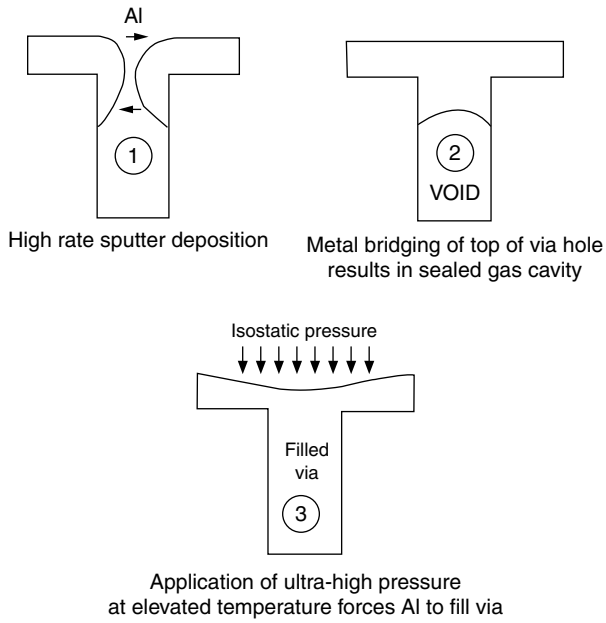


FIGURE 15.8 Sketch of the deposition process using high pressure extrusion. (1) High are rate PVD, (2) bridging or void formation, and (3) application of high gas pressure resulting in movement of the metal film down into the via. (From Rossnagel, S. M., *J. Vac. Sci. Technol.*, B16, (1998): 2585.)

throw distances, which require lower pressures. Most magnetrons will not operate below a fraction of a milliTorr without some means of enhancement, such as a hollow cathode electron source [32].

Long throw sputter deposition is also limited by an intrinsic asymmetry problem, shown in Figure 15.9. In the case of a sample position near the centerline of the system, the deposition is uniform

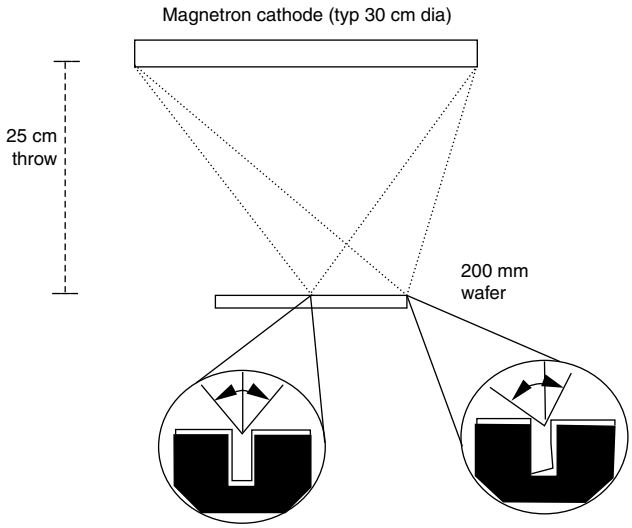


FIGURE 15.9 Sketch of center-to-edge asymmetry problem with long throw sputter deposition.

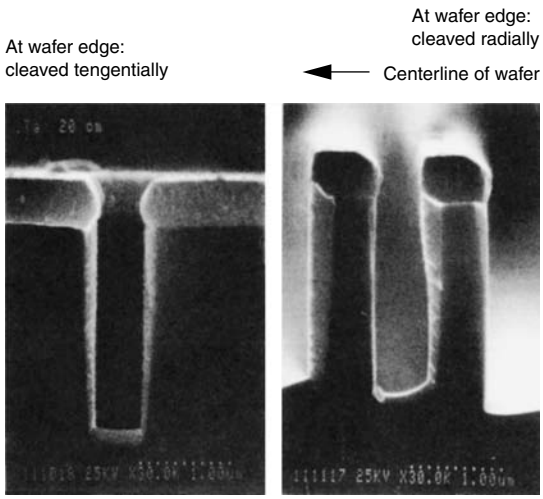


FIGURE 15.10 SEMs of vias located at the wafer edge showing the asymmetry of the depositing flux. (a) a trench cleaved in the direction perpendicular to the wafer edge, (b) a via cleaved in a direction parallel to the wafer edge. (From Mayo, A., S. Hamaguchi, J. H. Joo, and S. M. Rossnagel, *J. Vac. Sci. Technol.*, B15, (1997): 1788.)

from all angles up to the cutoff angle. However, near the edge of the wafer, the deposition is stronger from the inner regions of the cathode, resulting in a greater buildup on the outer sidewalls of deep features (Figure 15.10). The typical level of the asymmetry is 2–3×, and the ratio can be even higher at higher aspect ratio. Calculations as well as experiments have explored this problem at length, but in reality there are no simple solutions other than increased throw distance.

Similar geometrical arguments limit the extendibility of long throw deposition to the 300 mm wafer generation. Since the cathode size scales up linearly (from 30 to 45 cm diameter), to attain the same level of directionality would require increasing the throw distance by 50%, and at the same time reducing the pressure by 2×. In general, this technology does not scale well to 300 mm and is unlikely to be commercially available.

15.4.3.2 Collimated Sputtering

In a long mean free path deposition environment (mean free path ≫ throw distance), geometrical filtering of the deposition flux can also be obtained by placing a collimator between the target and the sample [34,35]. The collimator serves as a simple directional filter by simply collecting the atoms which impinge on its walls. This is shown schematically in Figure 15.11. The degree of filtering is simply a function of the aspect ratio of the collimator, where aspect ratio is defined as the thickness of the collimator divided by the diameter of a cell. The effect on the sputtered flux is shown in Figure 15.12, which shows the conventional emission distribution as a sphere centered about an impact site on the cathode surface. This sphere is the collection of all of the possible trajectories for the sputtered atoms. By increasing the aspect ratio of the collimator, the transmitted atomic distribution is shown as a cone centered about the surface normal. The higher the degree of collimation, the smaller the half-angle of the deposited cone of material.

The deposition rate obviously suffers during collimated sputter deposition. For each 1:1 increase in the aspect ratio of the collimator, the deposition rate falls by about 3× (Figure 15.13). Another way of picturing this is to consider the volume of the sphere in Figure 15.12, as compared to the volume of the cone of deposited material.

In collimated sputtering, it is generally not necessary to increase the throw distance significantly, other than the thickness of the collimator (typically 2–3 cm) and perhaps another centimeter or so to prevent direct shadowing of the collimator sidewalls on the wafer. This points the throw distance at about 8–9 cm,

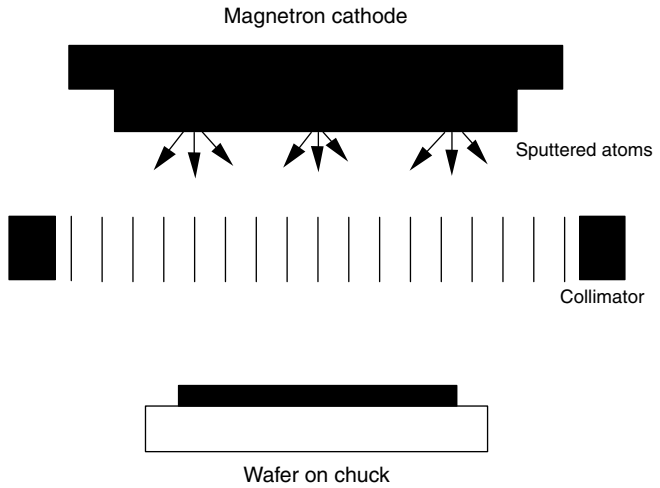
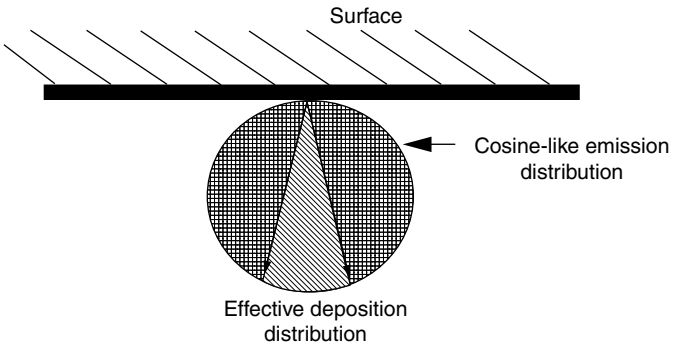


FIGURE 15.11 Collimated sputter deposition.

which requires an operating vacuum in the 0.5–1 mTorr range such that there are few in-flight gas phase collisions. This operating pressure is within range for most commercial, production scale magnetrons.

Collimator designs have undergone several iterations in the past 10 years. The original designs were machined from solid plates of Cu or Al: a close-packed hole pattern was machined by means of a numerically-controlled mill with sidewall thicknesses in the range of 40–50 thousands of an inch. The Cu or Al plates were water-cooled from the outer edge to prevent significant temperature buildup, although a center temperature of $>100°C$ could be obtained. Eventually to increase the transparency of the collimator, the holes were machined with a hexagonal, rather than circular diameter. Current (late 1990s) collimators use spotwelded arrays of thin sheet metal (typically Ti or stainless steel) which drop into



For a 2 cm-high collimator located 2 cm from cathode:

Aspect ratio	Emission width (degrees)
1:1	28 (i.e. +/–14)
2:1	14
3:1	11
4:1	7

FIGURE 15.12 Schematic of the emission distribution (shown as a sphere) and the subsequent filtering by a collimator.

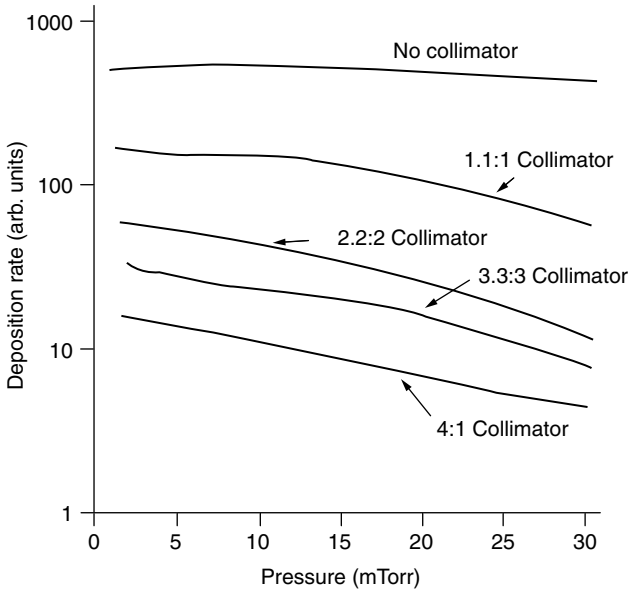


FIGURE 15.13 Deposition rate as a function of pressure through a variety of collimators.

uncooled housings and are held without fixtures or screws. The use of Ti is desirable for collimated Ti depositions because of the match between the thermal expansion coefficient of the collimator and the deposited films. However, these uncooled collimators can also reach temperatures of 500°C during deposition, particularly of TiN.

Collimator lifetime is usually limited by eventual closure of the collimator cell, rather than flaking. The lifetime is roughly on the order of half of the target lifetime, but this depends strongly on the material used. Originally collimators were recycled and recleaned, but this is rarely done today in production. A sheet-metal collimator for a 200 mm tool costs $600–2000.

15.4.3.3 Applications

Collimated sputtering was originally used for lift-off applications and then for filling trenches and vias [35]. Neither of these uses became practical commercially. In the semiconductor industry, collimated sputtering is mostly used for the deposition of Ti contact layers into the bottom of vias, and also for TiN diffusion barriers that are deposited prior to W-CVD. The bottom surface coverage (also known as "step coverage" is shown in Figure 15.14), as a function of the aspect ratio of the via. As expected, increasing the aspect ratio of the collimator results in better bottom coverage and the technique appears to be adequate for aspect ratios of up to 4:1. Higher aspect ratios require ever increasing collimator aspect ratios, and manufacturing applications have tended to stay at the 1:1 or 1.5:1 level for practical reasons.

A somewhat unexpected application for collimated sputtering was its application for near-conformal diffusion barriers or "liners" [36]. This was not anticipated because collimated sputtering was viewed as a directional deposition technology, which should preclude any significant deposition on the sidewalls of features. However, during the early stages of a deposition, the function of the collimator is to prevent or reduce the intrinsic build-out of the deposit in the upper corners of the via. This allows additional deposition down the sidewalls and towards the bottom corner, more than would be anticipated by conventional deposition alone. However, this is only the case for the early stages of deposition as it is applied to a near-conformal film. As the film thicken, shadowing will rapidly cutoff deposition farther down in the feature. For semi-conformal deposition, acceptable results can be obtained with a collimator which has an aspect ratio significantly less than that of the feature.

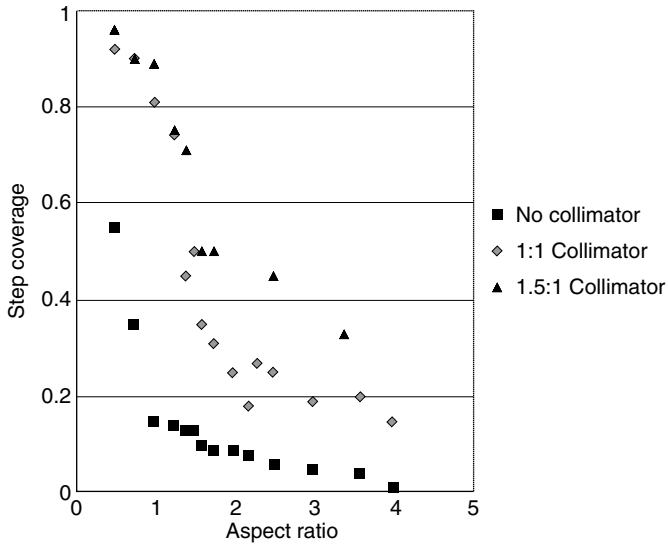


FIGURE 15.14 Bottom step coverage as a function of a via aspect ratio for uncollimated, 1:1 and 1.5:1 collimated sputter deposition. (From Rossnagel, S. M., *J. Vac. Sci. Technol.*, B16, (1998): 2585.)

From a practical point of view, these conformal films are less than ideal. There is a gradual taper from the upper corner to the bottom corner, and a significant crack at the bottom corner on all sides. In addition, the film structure on the sidewalls is distinctly columnar, resulting in unwanted pathways for diffusion across the film. One solution to this problem has been to use multiple-step collimation, such that the film is split into two layers which hopefully do not have identical crystallinity (42). An alternate is stuffing the films with extra N or O to help make it more inert.

The filling of moderate aspect ratio features is possible with collimated sputtering [20,34] but rarely practicesd. Filing has been demonstrated up to aspect ratios of 4:1, but this requires a collimator of equal to or greater aspect ratio. In addition, the sidewall deposits in the fill are underdense, similar to those observed with linear deposition, and this can provide poor resistance to electromigration. A second concern with filling applications is that it leaves behind a very think deposit on the field region of the wafer which must then be removed by means of chemical–mechanical polishing. This can add significant cost to the deposition process. An alternate scheme uses grazing-angle ion beam bombardment of the growing film (i.e., under the collimator) which can radically reduce the overburden and even reduce the amount of collimation required for good filling [37] This approach has not been extended to manufacturing, though, due to reliability concerns with broad beam ion sources.

15.4.4 Ionized Deposition

Physical sputtering is predominantly a neutral atom emission process: almost no ions are formed during the sputtering process, and even if an ion was formed, it would be held onto the surface by the electric field of the plasma sheath. Occasionally negative ions are formed in cases with very electronegative materials, but this is not an issue for almost all semiconductor materials [38]. The sputtered atoms are emitted with a wide range of angles, and since they are neutral, there is no other way than simple subtractive filtering to control their directionality.

In the late 1980s, there was significant work in the field of high-density plasma generation, primarily pointed towards etching applications. It became apparent during that work, though, that it was fairly easy to contaminate these plasma with metal atoms which had been sputtered evaporated or arced from the internal walls of the system. These metal atoms were readily ionized and could be used to diagnose the

etch plasmas. In general, though, this was considered a great nuisance since the metal ions would coat various insulating surfaces and windows in the high-density plasma tools and ruin their effectiveness.

It was not long, though, before people began to intentionally introduce metal into their plasmas as a way of intentionally depositing films, this time from primarily metal ions as opposed to metal atoms. The intrinsic advantage of metal ion deposition is that due to the nature of the plasma sheath, which is parallel to the sample surface, all of the ions are deposited at exactly normal incidence. Regardless of the original trajectory of the metal atoms (which might have been sputtered off some nearby surface at a random angle), the metal ion was accelerated across the sample sheath at 90° and the kinetic energy was set completely by the difference between the plasma potential and the wafer potential, both of which can generally be easily controlled. These two features are the great, intrinsic advantages of ionized deposition, or as will be described in this chapter ionized-physical vapor deposition (I-PVD).

The earliest work used both sputtered and evaporated sources [39,40] and a high-density plasma formed by electron-cyclotron resonance (ECR), which is driven by a microwave source at 2.45 GHz. It is necessary to shield the entry point for the microwaves from metal deposition, and this was done by placing the window behind a bend. The tool is operated by initiating an ECR discharge in Ar, and then starting the evaporation source (typically Al or Cu). The metal atoms can then be ionized by the density inert gas plasma, and at some point the argon can be removed by pumping and the plasma sustained completely by the evaporative source. Since the sample location is not in a direct line-of-sight to the evaporative source, only ions are deposited. This system was used for direct, ionized deposition of Cu into semiconductor features at aspect ratios of 4:1 [40]. Unfortunately, there was little enthusiasm for reintroducing evaporation as a semiconductor process technology on the manufacturing scale, so this ECR approach has been converted to a physical sputtering approach [41,42].

In parallel with the high-density microwave plasma work, there was also significant interest in inductively-coupled high-density rf plasmas, typically operated at 1.9–13.56 MHz. Coupling this with metal based plasmas, early work was done by Yamashita [43] and also by Barnes, Forster and Keller [44] which combined a metal sputtering source with a dense, inert gas inductively coupled plasma which was used for metal ionization. Later work was extended to high rate manufacturing-scale sources as well as measurements of the basic plasma properties [45–47]. The general inductively-coupled rf approach to I-PVD is shown in Figure 15.15. The magnetron cathode is conventional, i.e., it is the same cathode used for planar magnetron sputter deposition. In place of the collimator, a 1–3 turn rf coil is positioned

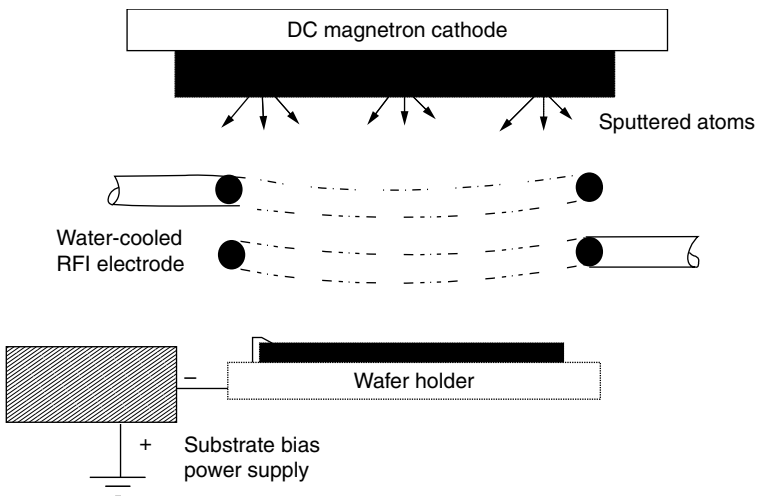


FIGURE 15.15 Experimental configuration for rf-based I-PVD. (From Rossnagel, S. M., *J. Vac. Sci. Technol.*, B16, (1998): 2585.)

approximately equidistant from the cathode and the sample, typically 3–4 cm from each. The coil diameter varies depending on the supplier and the group, and tends to be approximately the same diameter or slightly larger than the magnetron target diameter. It is important that the coil does not intercept the direct line-of-sight from the edge of the magnetron cathode to the sample, as this will result in shadowing near the wafer edge.

The function of the rf coil is to set up a dense, inductively coupled plasma in the background gas, which is typically Ar. The function of the magnetron is to sputter atoms into this discharge. At the sample, typically the sample potential is held by means of a clamp ring which, along with the sample pedestal, can be powered either rf or dc to a level of a few hundred watts at most. The wafer potential will typically be negative, and this will accelerate ions from the plasma, which has a positive plasma potential of a few volts, to the wafer. Using an rf bias overcomes problems with insulating wafer surfaces (or backsides), but results in an inability to measure actual sample currents.

The rf coil in early work was constructed of Cu tubing, and water cooling was supplied through the cooling to control the temperature. Coils of many sizes and dimensions were explored: varied numbers of turns, spiral coils, etc. The best results tend to come with a minimum number of turns (1–2) and the largest diameter tubing. These both tend to maximize the level of inductive coupling to the plasma, resulting in the highest plasma density. The rf coil, since it is exposed to the plasma, also develops a negative dc bias, typically of a few hundred volts. This can result in coil sputtering, even though the coil also receives a significant deposition flux from the plasma. By varying the tuning of the coil it is possible to operate the coil anywhere from a net deposition mode to a net etching mode. The former case allows the usage of a Cu coil for materials other than Cu, since the Cu will be buried under the depositing film. However, this film may also flake off over time resulting in particulate contamination. Operating the coil in the etch mode eliminates this problem, but requires that the coil be constructed from high purity material, since its atoms will be mixed in the discharge with the metal atoms from the magnetron, both going together to form the film. Commercial implementations of this approach have tended also to use non-water-cooled coils, since fabrication of water cooling in many materials (Ti, Ta, etc.) is nontrivial. The coil then runs hot, which may have negative effects on the film properties in some materials sets, particularly Al and Cu.

The relative ionization in the I-PVD rf system has been measured by using a gridded energy analyzer at the sample location. In place of a planar collector, the detector used a quartz crystal micro balance (Figure 15.16). This allows the detector to differentiate between inert gas ions and metal ions. The data from this type of detector is not directly related to the relative ionization level in the plasma because the presheath tends to pull metal ions to the sample. However, it is consistent with the depositing flux ionization ratio.

The relative ionization, as might be expected, tended to increase as chamber pressure was increased to the inductively coupled discharge (Figure 15.17). There was a slight different from Ar to Ne which may be attributable either to a higher electron temperature for Ne, or perhaps Penning ionization effects. However, it can be seen from the figure that relative ionization levels of 80–90% are possible. The maximum ionization was observed at pressures in the tens of mTorr range. At these pressures, the sputtered atoms tend to have many collisions in the gas phase, and as such, tend to stay the longest in the plasma region.

The relative ionization was also measured as a function of increasing rf power to the inductively coupled coil, as shown in Figure 15.18. In this case, the magnetron was operated at three different power levels, and these levels would scale approximately with the amount or number of metal atoms added to the discharge. At low metal fluxes (1 kW magnetron power), the ionization could be sustained at over 80%. However, as the metal flux was increased, the relative ionization was suppressed and could not be recovered by simply adding increased rf power. This was also observed in the ion current measured at the sample: increasing the metal flux to the plasma resulted in a reduction of sample ion current, consistent with a reduction in either plasma density or else electron temperature [48].

This general effect is quite troubling, because it is indicative of a less efficient plasma ionization process as the magnetron power is increased: something that is not favorable to high power scaling of the technology. Since the ion flux to a surface is directly proportional to the plasma density, which was

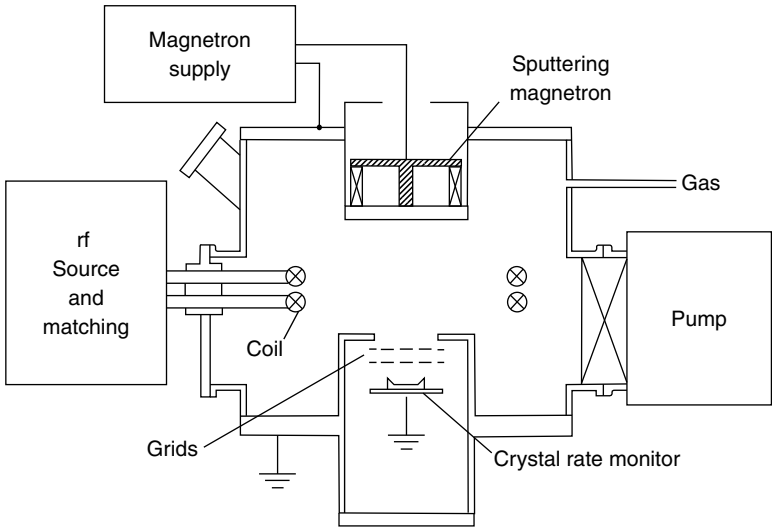


FIGURE 15.16 Retarding grid energy analyzer used to measure relative ionization of the deposition flux. (From Rossnagel, S. M. and Hopwood, J., *J. Vac. Sci. Technol.*, B12, 449, 1994.)

assumed to increase as the more-easily-ionized metal was introduced to the plasma, it was thought that the plasma temperature (actually the electron temperature) was cooling significantly. However, experiments to measure this did not show significant cooling and actually showed a slight decrease in plasma density, even though the added metal atoms should have been much easier to ionize than the background, inert gas atoms.

This paradox was resolved by the development of a quantitative discharge model by Dickson, Qian and Hopwood [49]. The model was able to predict the various contributions to the ionization and plasma density from both electron bombardment as well as Penning ionization processes. The model was useful at some predictions of the discharge properties, but in its original development was

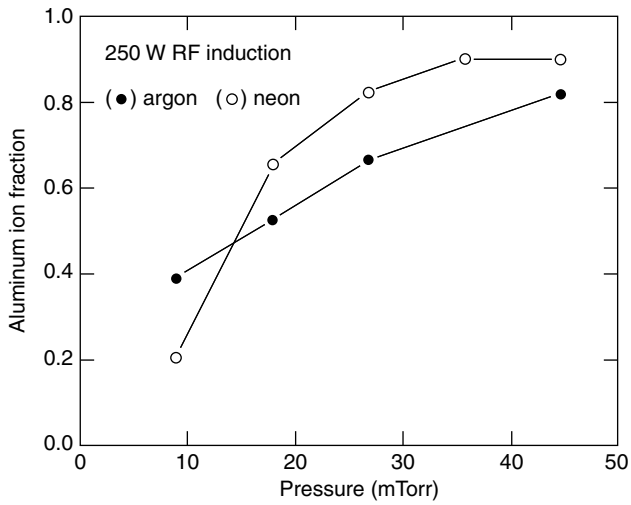


FIGURE 15.17 Relative ionization at the sample location for inductively coupled I-PVD as a function of increasing pressure. (From Rossnagel, S. M. and J. Hopwood, *J. Vac. Sci. Technol.*, B12, (1994): 449.)

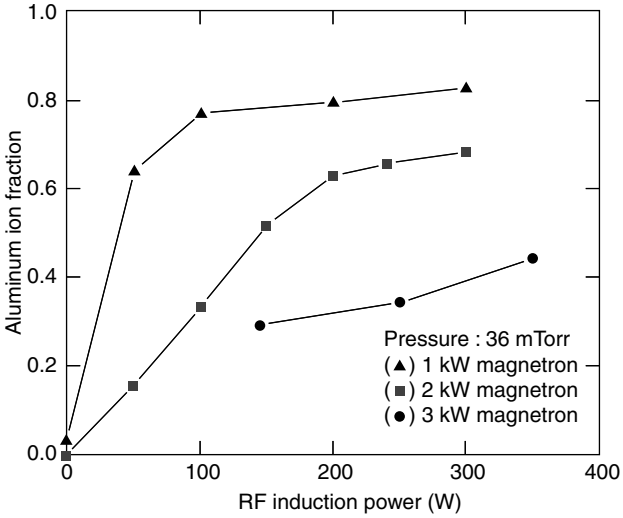


FIGURE 15.18 Relative ionization at the sample location for rf-based I-PVD as a function of chamber pressure for Ar and Ne. (From Rossnagel, S.M. and J. Hopwood, *J. Vac. Sci. Technol.*, B12, (1994): 449.)

unsuccessful at exploring significant increases in metal flux. However, they observed that if they allowed the gas density to decrease due to sputter-induced gas heating, the model correctly predicted the changes in discharge properties with added metal. This breakthrough was based on the earlier observations [17] of gas rarefaction and heating during magnetron sputtering. An example of this data is shown in Figure 15.19, which shows how the increased number of hot, sputtered metal atoms results in a measurable decrease in the gas density in the plasma region.

The model that emerges is as follows. As metal atoms are sputtered into the inductively-coupled, high-density inert gas plasma, some of these metal atoms transfer their kinetic energy to the background gas. The result of the gas heating (in an open chamber) is that the gas density declines slightly. As this density drops, the amount of time that a sputtered atom spends in the plasma region declines slightly because the

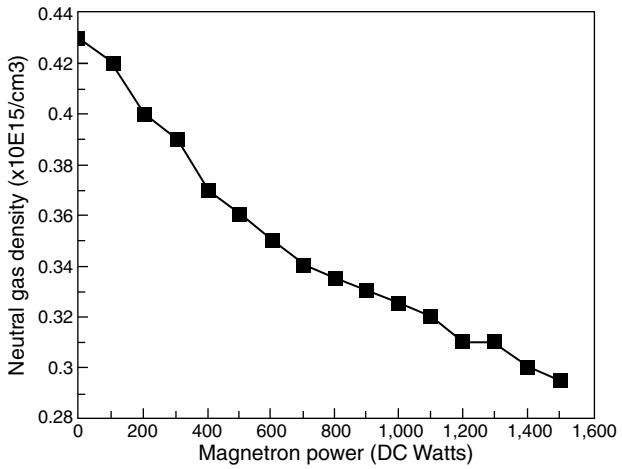


FIGURE 15.19 Change in neutral gas density at a fixed rf induction power of 1200 W as a function of increased magnetron discharge current. (From Rossnagel S. M., *J. Vac. Sci. Technol.*, B16, (1998): 2585.)

region is more transparent. And therefore, the likelihood of ionization of the metal atom is reduced. Adding more metal atoms to the discharge simply exaggerates the effect. It is almost as though the additional metal atoms result in a decreasing operating pressure, which according to the data of Figure 15.18, results in a less efficient ionization process. Several solutions exist, although the most obvious is to scale the inert gas density with the magnetron power such that the central gas density is fairly constant as more metal is introduced.

15.4.4.1 Applications

In many ways, I-PVD technology has very similar applications to the previously-described directional depositions based on filtering. By using the directional nature of the flux, depositions can be more easily made into deep, high aspect ratio features on semiconductor wafers. The three primary applications of this are bottom-contact layers, conformal coatings, and filling applications.

1. Contact layers. I-PVD techniques are ideal for projecting metal ions down to the bottom of a high aspect ratio via. In this application, the role of the deposited metal may be to make better electrical contact with some underlying contact or line, or else perhaps to deposit a metal for chemical incorporation into an under layer, such as a silicide of Ti or Co. The bottom step coverage (relative thickness as compared to the top, field thickness) as a function of aspect ratio is shown in Figure 15.20, comparing conventional PVD, collimated PVD and I-PVD [20]. As can be seen from the figure, high coverages of 50% or so can be observed at aspect ratios of 10:1. The step coverage declines slightly as the aspect ratio is increased due to small angle scattering across the sheath as well as small levels of ion temperature (fraction of eV) in the plasma. Actually, as a general rule of thumb, the relative ionization of the deposit can be inferred from the bottom step coverage at an aspect ratio of about 3:1. At smaller aspect ratios, there is a contribution to the bottom step coverage from the neutral deposition. At higher aspect ratios, scattering starts to become important.
2. Liners, diffusion barriers, adhesion layers and seed layers. One of the unexpected advantages of collimated sputtering was its usefulness at creating nearly-conformal coatings on the sides and

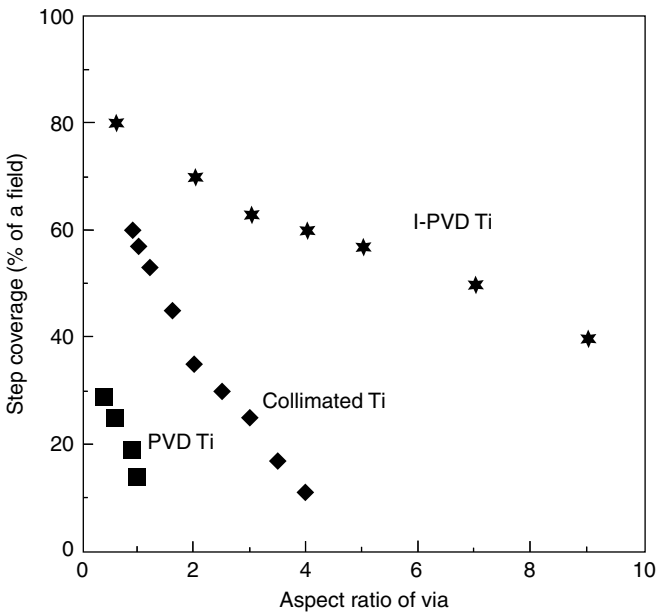


FIGURE 15.20 Bottom surface step coverage as a function of aspect ratio for conventional PVD, 1.5:1 collimated PVD, and I-PVD. (From Rossnagel S. M., *J. Vac. Sci. Technol.*, B16, 2585, 1998.)

bottoms of moderate aspect ratio features. This was unexpected because the collimated technology was intended to provide a more directional deposit, and conformality is generally inconsistent with high levels of directionality. I-PVD as a deposition technique is in many ways similar to collimation: the flux is mostly directional with a bit of scatter. So the initial expectation is that it would be similar to collimated sputtering.

It turns out that there is a distinct advantage of I-PVD over collimated sputtering for near-conformal films. This advantage is related to the ability to control the depositing ion's kinetic energy, usually by simply adding a negative bias to the sample. The result is that the ion energy can be increased sufficiently to cause physical sputtering, or resputtering of the deposited films [50]. When this occurs with a liner film, two advantages are seen (Figure 15.21). The first is that a small bevel forms at the top corners of the deposit, due simply to the fact that most materials have a slightly higher sputter yield at 45° or so than at normal incidence. This small bevel can taper back the overhang formation a little and tends to keep the via open. The second advantage occurs at the bottom of the feature, where atoms are sputtered from the bottom surface. These atoms are emitted with a cosine distribution and tend to end up depositing on the lower sidewall regions. The result is a slight thickening of the bottom corner deposition, which was exactly the place where the collimated deposition was weakest. This tapers in the bottom corners a little and makes subsequent filling of the via even easier.

This process of local resputtering can be used to tailor the conformality of the deposit to a great degree. The process has been easily extended to aspect ratios of 5:1, and may extend to 7:1. Unfortunately, as the aspect ratio increases, the number of atoms or ions incident on the aperture of the via does not increase, and the result is an ever-thinner coatings.

3. Filling of trenches and vias: the goal of filling for trenches and vias is to provide high-density, low resistivity metal with the desired micro-structure and orientation. Somewhat like collimation,

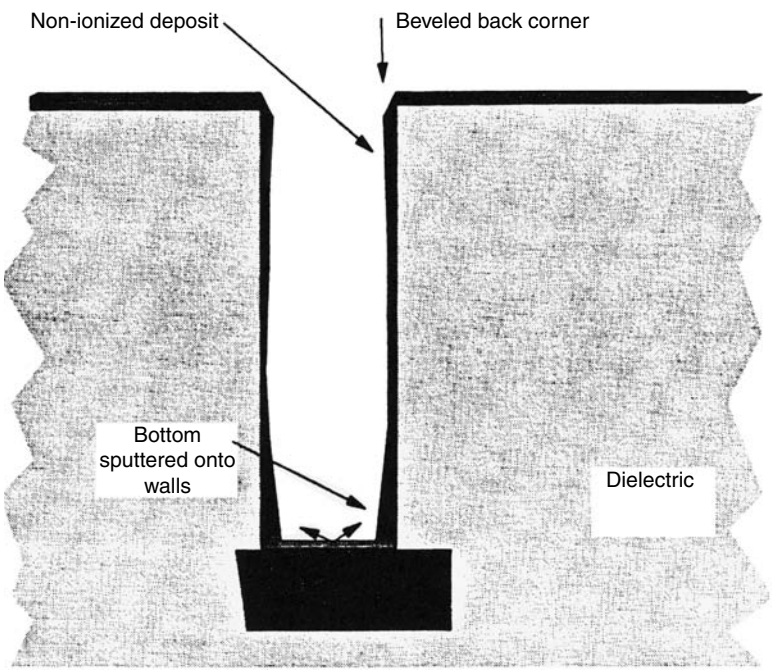


FIGURE 15.21 Sketch of I-PVD of a line showing the effects of resputtering of the deposited film.

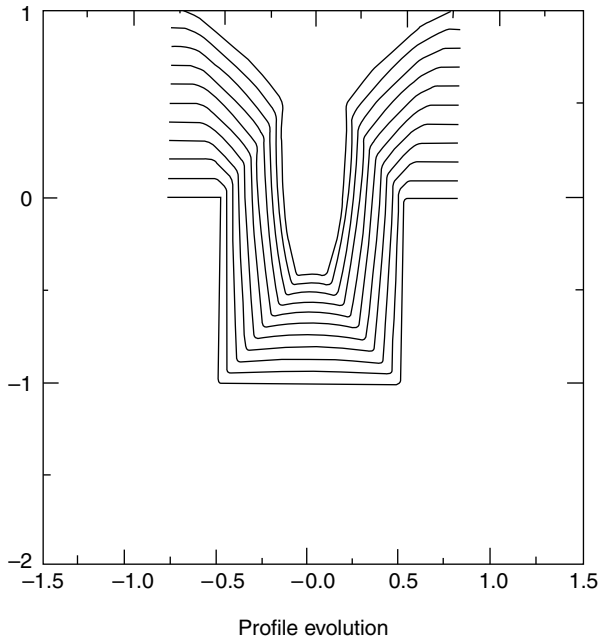


FIGURE 15.22 Computer simulation of the effect of resputtering at a feature aspect ratio of 1:1.

I-PVD alone will not turn out to be completely successful. Since the deposition on the sidewalls differs from the bottom deposit by being much more columnar and lower density, both collimated depositions or I-PVD depositions will need to be annealed at some fraction of the metal melting point to provide adequate recrystallization of the deposit.

I-PVD filling can be quantified by measuring the relative deposition rates on the sidewalls and bottom [48]. This data suggests that without additional sample, temperature filling is limited in a practical sense to an aspect ratio of about 2:1. Resputtering can be used at lower aspect ratio to taper back edges (Figure 15.22), but as the aspect ratio is increased, sputtering results in local redeposition across the trench, resulting in void formation (Figure 15.23).

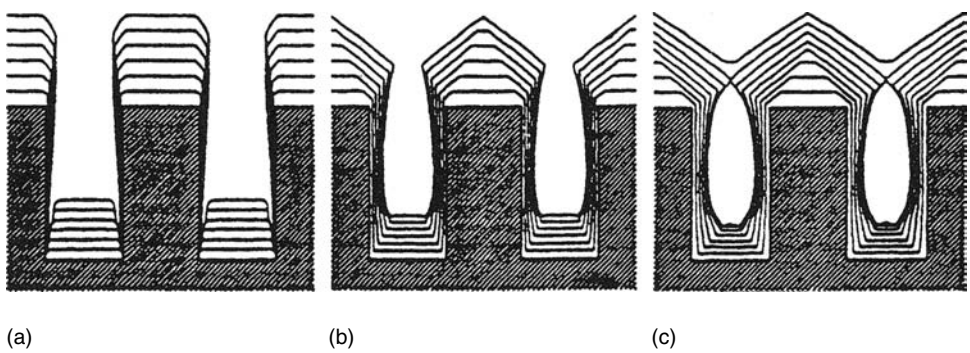


FIGURE 15.23 Simulations of the effect of increasing levels of resputtering *n* filling, using a 50% relative ionization. The sputter yield for (a)*n* was 0, for (b) was 0.4, and (c) was 1.0. (From Hamaguchi, S. and S. M. Rossnagel, *J. Vac. Sci. Technol.*, B13, 1995: 183.)

15.5 Summary, Future Directions

PVD technology has been one of the primary deposition techniques used for the manufacturing of semiconductor devices. Sputter deposition is used at almost all levels in the interconnect formation process, either in the form of thin films such as contact layers, diffusion barriers or seed layers, or for the primary conductors. The RIE-metallization process (Figure 15.1) was ideally suited for PVD, as the requirement was for smooth, planar films, which were continuous over small steps or lines on the sample surface. The other widespread usage of PVD in semiconductor manufacturing is for the Ti bottom layers and TiN diffusion barrier layers in vias, often deposited using collimated sputtering, which were then subsequently filled with W using a CVD process. These technologies have been adequate for semiconductor generations down to the 0.35 μm-width level.

As the feature size continues to decline and as the aspect ratio of the interconnect features continues to increase, it is obvious that conventional PVD or even filtered PVD (collimated or long throw) will not be adequate techniques due to the poor step coverage at high aspect ratio (AR) and the overhang formation. This has led to the I-PVD techniques, as well as continuing work in CVD technologies and electroplating as eventual replacements for PVD.

CVD technologies have some additional degrees of flexibility that are not present with PVD. For example, depending on the sample temperature, the gas dynamics, and even the substrate surface composition, CVD films can be deposited with almost 100% conformality at very high aspect ratio. In addition, CVD film composition and purity can also be controlled by adjustments to the process gas or the sample temperature. Most of the materials relevant to interconnect technology: Al, Cu, Ti, TiN, W, and TaN can be deposited by CVD, generally in a temperature and environment which is compatible with wafer processing.

There are several challenges, though, with CVD technology which has limited its widespread usage for all materials aside from W-CVD. For example, Al CVD has been demonstrated but only for pure Al, whereas the semiconductor industry requires small levels of Cu doping in the films for electromigration resistance. It is possible to get around this by depositing a thin layer of CVD-Al, followed by a PVD AlCu, which after a subsequent temperature cycle the PVD material donates some of its Cu to the CVD-Al layer. This is a variation on the two-step reflow process described above, and is available commercially. The conformality of the CVD-Al is useful at high aspect ratio as a seed layer for the subsequent, conventional PVD reflow deposition. A similar technique has been developed for CVD-Cu: a thin seed layer of CVD-Cu followed by an elevated temperature Cu reflow deposition using conventional PVD.

CVD techniques for Ti and TiN are often constrained by the presence of impurities, generally from the precursors. CVD-Ti, which is typically deposited from $TiCl_4$, can have trace Cl levels in the deposited film. The TiN, deposited from either a TDMAT or TDEAT precursor at temperatures i
n the 400°C range can have significant levels of C or O in the films, which can partly be alleviated by a nitrogen plasma treatment. The resulting films, though, have significantly higher resistivity than PVD films. In general, CVD technologies are characterized by increased cost and complexity, as well as subtle chemical problems, compared to equivalent PVD techniques such as collimation or I-PVD.

Electroplating has also become viable as a potential interconnect deposition technique for Cu applications for filling high aspect ratio vias and trenches. (Electroplating is not practical for Al deposition.) Since Cu rapidly interdiffuses into Si forming deep traps, Cu is unlikely to be used for the first layer of metallization (M-0); those vias which contact the wafer surface will most likely remain as CVD-W. However, in the upper interconnect layers, Cu's lower resistivity compared to AlCu makes it an ideal candidate for higher speed lines and vias. Since Cu can diffuse into the dielectric (although at a much slower rate than into Si), it is necessary to use a refractory diffusion barrier, such as TiN, TaN, or W. It may also be necessary to use an adhesion layer, such as Ti, Ta or Cr, as Cu does not adhere well to many surfaces, such as silicon dioxide. Finally, to reliably plate onto the sample surface, a seed layer of Cu is needed.

The conformal nature of the diffusion barrier, adhesion layer (if used) and seed layer can be addressed with either I-PVD films or CVD techniques. Since the aspect ratio of mid- and upper-level lines and vias is not likely to exceed 5:1, it is probable that I-PVD techniques will become widely used for these films, due to lower cost and the intrinsic simplicity of the I-PVD tools. Commercial tools have recently been announced, and are just beginning to reach the manufacturing floor. It is also apparent that I-PVD techniques will completely replace collimated PVD at the 300 mm wafer generation. The efficiency and cost of I-PVD deposition is significantly better than collimation.

I-PVD techniques are likely to be widely used for contact, barrier, adhesion and seed layers, but it is less clear if I-PVD will be used as a practical, cost-effective primary conductor. While there has been hesitation to consider electroplating as a legitimate semiconductor manufacturing process, there may be significant cost and performance advantages for electroplated films. While I-PVD filling techniques have been demonstrated at modest aspect ratios, significant work remains to be done to determine whether it can become a cost-effective technique for Cu. Since PVD-based manufacturing tools will continue to be needed for the various barrier and seed layers, I-PVD Cu filling may prove cost effective in that it resides on the same tool base and can easily integrated with the prior layers, as opposed to a separate plating platform.

The nearing transition to 300 mm wafer diameter provides a number of challenges to the interconnect technology area. Aside from the obvious increase in the physical size of the deposition and etching systems, such topics as increased uniformity or reduced particulate counts will need to be addressed. PVD systems are well-placed to make the transition to 300 mm. The scaling of the magnetron cathode is a complex, but fairly-well understood technology, and much larger cathodes have been built for other industries such as flat panel displays or glass coating. The increase in cathode diameter, as mentioned above, will scale as roughly 1.5× the wafer diameter, although the cathode-to-sample distance will remain constant at 4–10 cm. This means that the tools will become effectively even more two-dimensional and edge effects and the presence of chamber walls or tooling will be less important.

For directional sputter deposition, it is unlikely that either collimated sputter deposition or long throw deposition techniques will be widely used at 300 mm, and may not even be commercially developed. The I-PVD technology, based on the inductivelycoupled rf-coil approach, has been demonstrated at 300 mm sample diameter, and is most likely to be the dominant process used. Depending on the approach used: either rf coil or Faraday shielded coil, the dimensions of the tools will be modified slightly. This is necessary to account for the presence of a significant sputtered flux from the coil in the etched-coil mode, or the capturing effect of the Faraday shield, reducing the edge plasma and metal density near the shield. However, there is significant latitude in the design of the magnetron cathode emission pattern to compensate for these two issues.

References

1. Licata, T. J., E. G. Colgan, J. M. E. Harper, and S. E. Luce. *IBM J. Res. Dev.* 39 (1995): 369.
2. Kaanta, C. W., S. G. Bombadier, W. J. Cote, W. Hill, G. Kerszykowski, H. S. Landis, D. J. Poindexter, et al. *Proceedings of the 8th International VLSI Multilevel Interconnection Conference*. 144. Santa Clara, CA, 1991.
3. Zalm, P. *Surf. Interface Anal.* 11 (1988): 1.
4. Hofer, W. O. In *Sputtering by Particle Bombardment III*, edited by R. Behrisch, and K. Wittmaack, Berlin: Springer, 1991, chap. 2.
5. Anderson, H. H., and H. L. Bay. In *Sputtering by Particle Bombardment I*, edited by R. Behrisch, 145. Berlin: Springer, 1981.
6. Ruzic, D. In *Handbook of Plasma Processing Technology*, edited by S. M. Rossnagel, J. J. Cuomo, and W. D. Westwood, Park Ridge, NJ: Noyes Publication, 1989.
7. Ruzic, D. N., P. C. Smith, and R. B. Turkot Jr., *J. Nucl. Mater.* (1996).

8. Rossnagel, S. M. In *Handbook of Vacuum Science and Technology*, edited by D. Hoffman, B. Singh, and J. E. Thomas, 609. Orlando, FL: Academic Press, 1997.
9. Brauer, G., D. Hasselkamp, W. Kruger, and A. Scharmann. *Nucl. Instrum. Methods* B12 (1985): 458.
10. Wehner, G. K., and D. Rosenberg. *J. Appl. Phys.* 31 (1960): 177.
11. Fan, J. S., R. S. Bailey, and C. E. Wickersham Jr. SEMICON- China, November 1997 (unpublished).
12. Urbassek, H. M., and D. Sibold. *J. Vac. Sci. Technol. A* 11 (1993): 676.
13. Motohiro, T., and Y. Taga. *Thin Solid Films* 112 (1984): 161.
14. Somekh, R. *J. Vac. Sci. Technol.* A2 (1984): 1285.
15. Helmer, J. C., and C. E. Wickersham. *J. Vac. Sci. Technol.* A4 (1986): 408.
16. Rossnagel, S. M. *J. Vac. Sci. Technol.* A6 (1988): 3049.
17. Rossnagel, S. M. *J. Vac. Sci. Technol.* A6 (1988): 19.
18. Rossnagel, S. M., and H. R. Kaufman. *J. Vac. Sci. Technol.* A6 (1988): 223; Rossnagel, S. M., and H. R. Kaufman. *J. Vac. Sci. Technol.* A5 (1987): 2276.
19. Class, W. H. *Thin Solid Films* 107 (1983): 379.
20. Rossnagel, S. M. *J. Vac. Sci. Technol.* B16 (1998): 2585.
21. Logan, J. S. In *Handbook of Plasma Processing Technology*, edited by S. M. Rossnagel, J. J. Cuomo, and W. D. Westwood, 140. Park Ridge, NJ: Noyes Publications, 1989.
22. Restaino, D., S. Chiang, and G. Odlum. Update, *Applied Materials*. Vol. 4, 2. Santa Clara, CA, 1997.
23. Powell, R. A., and S. M. Rossnagel. *PVD for Microelectronics*. Boston, NY: Academic Press, 1998.
24. Inoue, M., K. Hashizumi, and H. Tsuchikawa. *J. Vac. Sci. Technol.* A6 (1988): 1636.
25. "Warm Al" process, *Applied Materials*, Santa Clara, CA, also Chang, B., et al. *Proceedings of VLSI Multilevel Integration Conference*. 389. Santa Clara, CA, 1997.
26. Pramanik, D., and A. N. Saxena. *Solid State Technol.* 33 (1990): 73.
27. Gardner, D. S., and D. B. Fraser. In *Proceedings of the 12th International VLSI Multilevel Metallization Conference*. 287. Santa Clara, CA, 1995.
28. Dixit, G., W. Y. Hsu, K. H. Mamamoto, M. K. Jain, L. M. Ting, R. H. Havemann, C. D. Dobson, et al. *Semicond. Int.* Aug (1995): 79 (note: the ElectroTech Company was acquired by Trikon (formerly PMT) in 1996).
29. Broughton, J., C. Backhouse, M. Brett, S. Dew, and G. Este. *Proceedings of the 12th International VLSI Multilevel Interconnection Conference*. 201. Santa Clara, CA, 1995.
30. Macwawa, K., K. Mori, A. Ohaski, M. Hirayama, C. D. Dobson, A. I. Jeffries, P. Rich, D. Butler, N. Rimmer, and A. McGeown. In *Advanced Metallization and Interconnect Systems for ULSI Applications in 1995*, edited by R. Ellwanger, and S.-Q. Wang, 341. Pittsburgh, PA: Material Research Society, 1996.
31. Motegi, N., Y. Kshimoto, K. Nagatani, S. Takahashi, T. Kondo, Y. Mizusawa, and I. Nakayama. *J. Vac. Sci. Technol.* B13 (1995): 1006.
32. Cuomo, J. J., and S. M. Rossnagel. *J. Vac. Sci. Technol.* A4 (1986): 393.
33. Mayo, A., S. Hamaguchi, J. H. Joo, and S. M. Rossnagel. *J. Vac. Sci. Technol.* B15 (1997): 1788.
34. Rossnagel, S. M., D. Mikalsen, H. Kinoshita, and J. J. Cuomo. *J. Vac. Sci. Technol.* A9 (1991): 261.
35. Mikalsen, D. J., and S. M. Rossnagel. U.S. Patent 4,824,544, Apr. 25, 1989.
36. Joshi, R. V., and S. Brodsky. *Proceedings of the 9th International VLSI Multilevel Interconnection Conference*. 253. Santa Clara, CA, 1992.
37. Rossnagel, S. M., and R. Sward. *J. Vac. Sci. Technol.* A13 (1995): 156.
38. Cuomo, J. J., R. J. Gambino, J. M. E. Harper, J. D. Kuptsis, and J. C. Weber. *J. Vac. Sci. Technol.* 15 (1978): 281.
39. Kidd, P. *J. Vac. Sci. Technol.* A9 (1991): 466.
40. Holber, W. M., J. S. Logan, H. Grabarz, J. T. C. Yeh, J. B. O. Caughman, A. Sugarman, and F. Turene. *J. Vac. Sci. Technol.* A11 (1903): 1993.
41. Gorbatkin, S. M., D. B. Poker, R. L. Rhodes, C. Doughty, L. A. Berry, and S. M. Rossnagel. *J. Vac. Sci. Technol.* B13 (1853): 1996.
42. *High Density Plasma Vapor Deposition*, AX4X50, Woburn, MA: ASTex Corporation.
43. Yamashita, M. *J. Vac. Sci. Technol.* A7 (1989): 151–8.
44. Barnes, M. S., J. C. Forster, and J. H. Keller. U.S. Patent 5,178,739, Jan. 12, 1993.

45. Rossnagel, S. M., and J. Hopwood. *Appl. Phys. Lett.* 63 (1993): 3285–7.
46. Rossnagel, S. M., and J. Hopwood. *J. Vac. Sci. Technol.* B12 (1994): 449.
47. Rossnagel, S. M. *Semicond. Int.* Feb (1996): 99.
48. Nichols, C. A., S. M. Rossnagel, and S. Hamaguchi. *J. Vac. Sci. Technol.* B14 (1996): 3270.
49. Dickson, M., F. Qian, and J. Hopwood. *J. Vac. Sci. Technol.* A15 (1997): 340–4.
50. Hamaguchi, S., and S. M. Rossnagel. *J. Vac. Sci. Technol.* B14 (1996): 2603.
51. Hamaguchi, S., and S. M. Rossnagel. *J. Vac. Sci. Technol.* B13 (1995): 183.

16

Damascene Copper Electroplating

16.1 Introduction **16**-1
Damascene Process • The Need for Copper • Alternatives to Electrodeposition for Damascene Applications • Cu Electroplating Evolution and Applications • Cu Deposition for Damascene Applications

16.2 Fundamentals of Electroplating **16**-8
Kinetics • Mass Transfer • Geometry Effects on Local Kinetics

16.3 Damascene Cu Electroplating Chemistry **16**-13
Electrolytes • Organic Additives • Anodes and Anode Films

16.4 Damascene Film Deposition **16**-19
Feature Fill and Leveling Capability and Mechanism • Fill Evolution Behavior • Fill Response to Plating Chemistry • Leveling after Fill • Fill Behavior in Alternate Electrolytes • Current Waveform and Mass Transfer Impact on Filling • Summary of Additive Behavior during Filling • Deposit Planarity and Surface Roughness • Thickness Distribution • Terminal Effect • Field Effects • Mass Transfer • Metallurgy and Reliability

16.5 Modeling Capabilities **16**-38

16.6 Process Integration **16**-40
Feature Profile and Seed Interactions

16.7 Process Control Approaches **16**-41

Acknowledgments **16**-42

References **16**-42

Jonathan Reid
Novellus, Inc.

16.1 Introduction

16.1.1 Damascene Process

Formation of interconnect circuitry for integrated circuits (IC) or other electronic components is normally accomplished through either subtractive or additive processing. In subtractive processing, a continuous layer of metal is deposited on a dielectric substrate. The interconnect circuit pattern is then generated by etching away undesired metal, usually as defined by a photoimaged resist pattern. Interconnects for IC's have traditionally been formed by subtractive dry etching of blanket aluminum alloy films to produce the desired circuit lines as defined by a photoresist pattern. In a standard additive process, a circuit pattern is generated in photoresist laminated onto a thin metal seed covering a dielectric

surface. Metal is next deposited, usually by electroplating, on the seed surface which is defined by the resist pattern to form full-thickness metal lines. Finally, the resist and seed layer are removed leaving only the metal circuit pattern. Additive processes have traditionally been used in high-end printed circuit board and flexible circuitry applications.

A unique variety of additive processing, known commonly as the damascene process, was introduced by IBM [1–3] in the 1990s as a means for the formation of copper IC interconnects. Using this process, a line or via pattern is defined by photo-processing and this pattern is etched into a dielectric layer. Following etching of the dielectric, the photoresist is removed and a copper diffusion barrier, such as tantalum, and a copper seed layer are sequentially deposited on all surfaces of the dielectric. The seed layer provides the conductivity across the wafer necessary for the electroplating process, as well as a surface upon which nucleation of the electroplated film can begin. Copper is then electroplated to form the desired circuitry by filling the seeded recesses in the dielectric [4]. Using this process, however, copper is also electroplated on the entire surface of the wafer as a continuous film. This excess copper is removed in a subsequent chemical mechanical polishing (CMP) step to define the circuit pattern. A variation of the damascene process, known as the dual damascene process, has been widely implemented as the most cost-effective means of producing copper interconnect circuitry. Using dual damascene processing, a via layer and an interconnect line layer are sequentially photoimaged and etched prior to seeding and plating. This eliminates several process steps compared to either two single damascene process steps or traditional subtractive aluminum processing with tungsten vias. The dual damascene plating sequence is illustrated in Figure 16.1.

16.1.2 The Need for Copper

The cost savings associated with the copper dual damascene process sequence may be significant in some applications, however, the transition from aluminum to copper interconnects has been driven primarily by three technical reasons [1–3]. First, the resistivity advantage of copper over aluminum has significantly reduced RC delay times in advanced logic devices, as well as allowed for smaller line dimensions

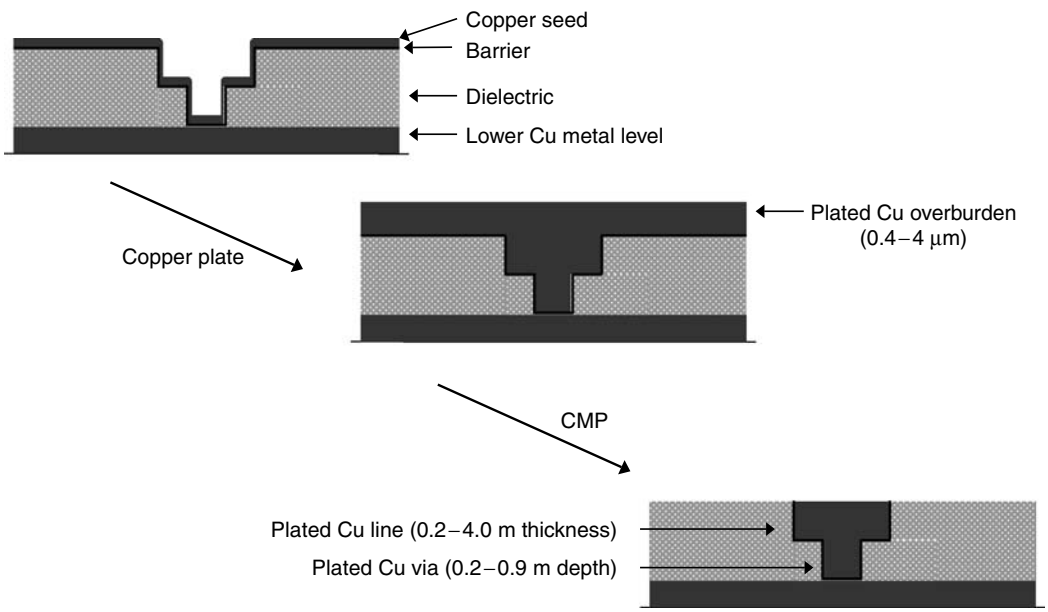


FIGURE 16.1 Damascene line formation sequence.

without increased heat generation. Second, as circuit dimensions decrease and line current densities increase the Electromigration (EM) life advantage of copper relative to aluminum becomes critical in ensuring device life. Finally, the extendibility of aluminum etch to circuit dimensions below 0.10 μm has been difficult, while damascene processing of copper is easily extendable below these dimension. Taken together, the advantages of copper interconnects result in interconnect solutions with flexibility to achieve improved EM life, higher device speed, less heat generation, and fewer interconnect metal levels depending on specific requirements. As a result of these advantages, nearly all high performance logic devices have made the transition to copper, as of 2004 [5].

16.1.3 Alternatives to Electrodeposition for Damascene Applications

Electrodeposition on a physical vapor deposition (PVD) seed layer to form dual damascene circuits was not initially recognized as a preferred means of copper interconnect formation. Early development of copper interconnect formation concentrated on chemical vapor deposition (CVD) [6,7], electroless copper deposition [8,9], and PVD [10] methods. In some cases, these methods were combined with each other or with copper reflow processes [11] as a means for deposition of seed layers or bulk copper films. Nearly all studies concluded that optimum adhesion of copper to barrier films was obtained using PVD copper deposition, thus necessitating the use of PVD copper as an initial seed layer. For bulk copper deposition, electroless, and CVD processes available resulted in a conformal film of copper within features and on the wafer surface. As shown in Figure 16.2, this can lead to a seam or void in the center of copper lines if the dielectric etch profile is re-entrant or near vertical. Such seams present reliability concerns in a metal such as copper, which can undergo EM. As also shown in Figure 16.2, PVD processes result in large voids within features when used for bulk copper deposition owing to the geometric limitation of metal flux into the feature relative to the flux on the field. While very high pressure and temperature have been shown to reflow copper into features, this approach has not been put into practice.

By providing a continuous film across the wafer surface and within features, PVD copper seed meets the basic requirement of electrical continuity needed for subsequent copper electrodeposition. Subsequent electrodeposition processes would be expected to yield conformal film deposition and seam voids as observed for CVD or electroless copper plating, however, it was recognized that copper electrodeposition using certain organic additives could result in accelerated copper growth from the base of high aspect ratio (AR) features resulting in void-free filling as shown in Figure 16.2. This capability of void-free filling, along with the near bulk resistivity value of the plated Cu and the relatively low cost of electrodeposition, led to its selection for interconnect formation.

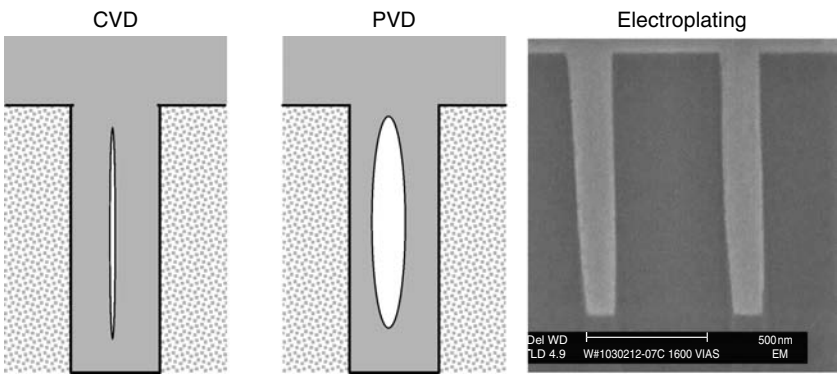


FIGURE 16.2 Profiles of copper deposits in high aspect ratio features formed using CVD, PVD, and electroplating processes.

16.1.4 Cu Electroplating Evolution and Applications

Electrodeposition of metal is performed by immersing a conductive surface in a solution containing ions of the metal to be deposited. The surface is electrically connected to an external power supply and current is passed across and through the surface to be plated causing reaction of the metal ions (M^{z+}) of charge (z) with electrons (e^-) to form metal (M):

$$M^{z+} + ze^- = M. \tag{16.1}$$

Copper electrodeposition from sulfate solutions was first investigated in the early 1800s [12]. Important industrial applications since that time have included flatware plating, wire, and foil formation, and printed circuit and connector formation. Concurrently, the relatively simple reaction kinetics of cupric ion reduction in sulfate solutions resulted in its use as the standard chemistry for most early fundamental electrochemical studies. Recently, interest in copper plating as a method of forming IC interconnects has increased. The evolution of activities related to copper electrodeposition is summarized in Figure 16.3.

16.1.5 Cu Deposition for Damascene Applications

In the case of electrodeposition of copper onto a silicon wafer, the wafer is typically coated with a 50–300 Å thick Tantalum based diffusion barrier and a 300–2000 Å thick copper seed layer prior to plating. The seed layer should have acceptable adhesion to the barrier layer, provide adequate conductivity across the surface of the wafer, and be continuous with full coverage in high aspect ratio features to be filled during the plating process. As illustrated in Figure 16.4, the seeded wafer is immersed in a solution containing cupric ions, sulfuric acid, chloride ion and proprietary additives. Electrical contact is made to the seed layer and current is passed such that the reaction $Cu^{2+} + 2e^- \rightarrow Cu(O)$ occurs at the wafer surface. The wafer is referred to as the cathode. Another electrically active surface, known as the anode, is present in the conductive solution to complete the electrical circuit. At the anode, an oxidation reaction occurs which balances the current flow at the cathode, thus maintaining electrical neutrality in the solution. When a copper anode is used during copper plating, all cupric ions removed from solution by deposition onto the wafer are replaced by dissolution of the anode.

In the absence of any secondary reaction, the current delivered to a conductive surface during electroplating is directly proportional to the quantity of metal deposited (Faraday's law of electrolysis).

$$W = ItA_w/nF \tag{16.2}$$

where W, weight of deposit in grams; t, time in seconds; I, current in Amperes; F, Faraday constant (96,500 C/eq); n, number of electrons transferred per atom deposited; A_w, atomic weight.

Using this relationship, the mass of Cu deposited can be readily controlled through variations of plating current and time.

With no applied potential and no imposed current flow across the interface between a metal and a solution, an equilibrium potential exists between the two. This potential is often known as the rest potential. Once potential is shifted from the equilibrium potential by an external power source, a current will be driven across the interface. Under conditions typical of most plating processes, this current density (I) is approximated by an exponential relationship known as the Tafel equation.

$$I = i_0[e^{-\alpha nF\eta/RT}] \tag{16.3}$$

where i_0, exchange current density; α, charge transfer coefficient; η, applied potential (V); T, absolute temperature; R, gas constant.

Figure 16.5 shows a current–potential curve typical of a copper deposition process. As potential applied to the wafer is scanned from the equilibrium potential to more negative values, the current

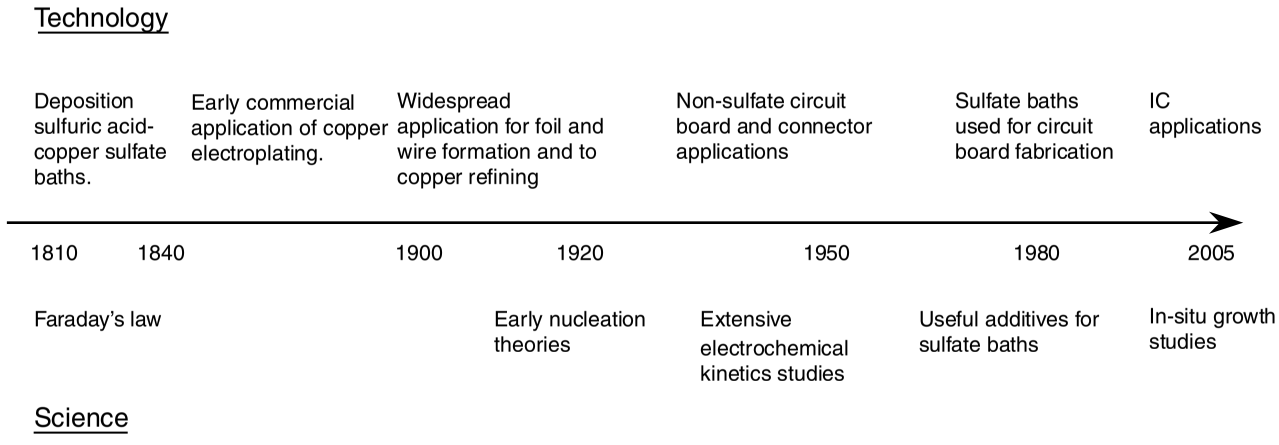


FIGURE 16.3 Evolution of copper electroplating.

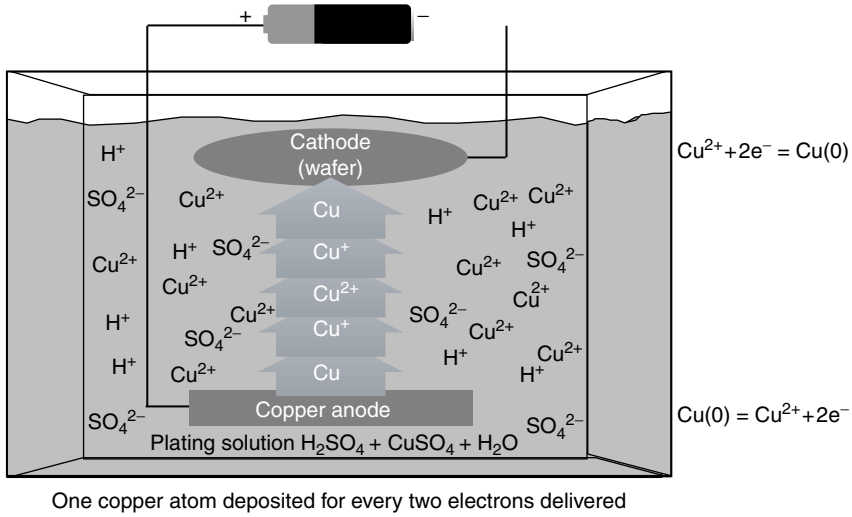


FIGURE 16.4 Schematic diagram of copper electroplating cell.

increases in an exponential manner (Tafel region) where the overall deposition rate is determined largely by charge transfer or reaction rate kinetics at the cathode. This strong dependence of current upon potential results in the need for plating cell designs which yield uniform potentials across the wafer surface.

As the potential continues to increase, mass transfer limitation on current gradually becomes dominant and a limiting current plateau is reached. At the limiting current species reacting at the cathode (Cu^{2+}) no longer reach the interface at a rate sufficient to sustain the rate of reaction possible at a high applied potential. As a general rule, plating processes are operated at currents no greater than 30%–50% of the limiting current in order to avoid undesirable deposit characteristics.

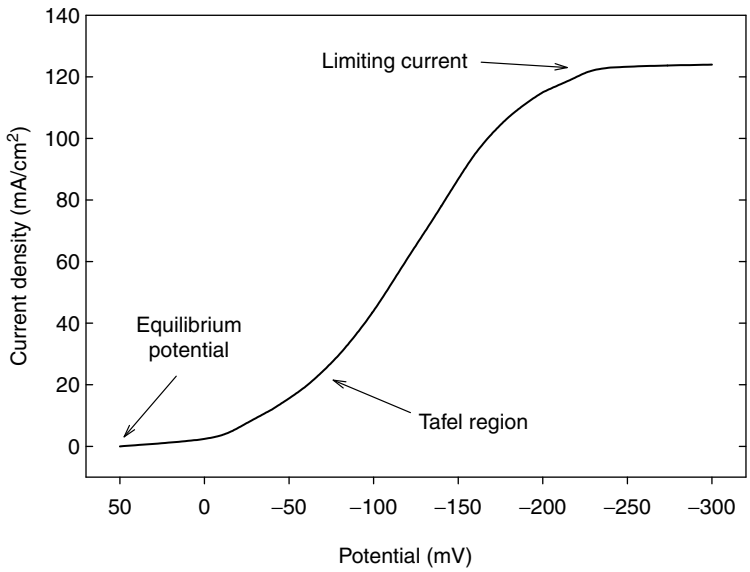


FIGURE 16.5 Current–potential relation for Cu electroplating.

To ensure that the rate of mass transfer of electro-active species to the interface is large compared to the reaction rate, and uniform across the wafer surface, the rates of cupric ion migration, diffusion, and convection must be understood and controlled. During wafer electroplating, convection must be accurately controlled while diffusion and migration are characteristic of a process but not normally controlled. Mass transfer due to convection can vary from stagnant to laminar or turbulent flow depending on hardware configuration. Convection includes impinging flow caused by solution pumping, flows due to substrate movement, ultrasonics, and flows resulting from density variations.

Electroplating can be carried out using a constant current, a constant voltage, or waveforms involving variable current or voltage. Using a constant current, accurate control of the mass of deposited metal is most easily obtained based on Faraday's law. Plating at a constant voltage and using variable waveforms requires more complex equipment and control, but can be useful in tailoring specific thickness distributions and film properties. A constant voltage also allows deposition at a uniform current density while plated surface area changes.

Copper plating bath components include sulfuric acid, copper sulfate, chloride ion, and trace organic additives [12,13]. Copper sulfate is added to baths to maintain the concentration of cupric ions in the range of 10–70 g/L. Low concentrations of cupric ion are utilized to obtain optimum thickness distribution control [14]. High cupric ion concentrations allow generally higher plating currents (deposition rate) and thus can allow increased throughput. Sulfuric acid concentrations up to 300 g/L can be utilized. High acid concentrations are often chosen to provide high solution conductivity and thus reduce electric field variability in the bath. A uniform electric field is essential to obtain a uniform current density and deposition rate [15]. Low acid concentrations uniformly increase bath resistance between the wafer and the anode and thus reduce the relative impact of seed resistance on current flow across the wafer. This reduces center-to-edge deposition rate variability on thin seeds which otherwise tend to plate more rapidly near the wafer edge where electrical contact is made [16,17]. Chloride ions are present in the bath at concentrations in the range of 30–100 mg/L.

Complex mixtures of organic chemicals are added to copper plating baths to influence deposit metallurgy and deposit thickness distribution. A typical additive mixture contains one or more chemicals which act to increase the current at a given voltage (accelerators), and at least one other class of molecules that act to reduce current at a given voltage (suppressers). Together, and at proper concentrations, these additives largely determine the desired filling behavior, grain structure, and purity of the deposited copper and as a result control ductility, hardness, surface roughness, stress, and tensile strength.

Primary responses to be considered during process optimization for copper electroplating of IC interconnects include filling performance, within die uniformity, within wafer uniformity, grain size, copper electromigration behavior, defects in the deposited film, and defects in the circuitry following CMP. Variables that will impact these responses include concentrations and types of accelerators, suppressors, and leveler, concentrations of chloride ion, copper, and acid, and other parameters such as flow, temperature, current density, and current waveform [18,19]. Achievement of robust filling performance required for a given product usually determines the general parameter space over which all responses must be investigated by process window experiments.

In order to fully fill high aspect ratio features with copper during electroplating, it is essential that the plating process proceed at an accelerated rate within the features compared to the rate on the surface of the wafer. This super-filling behavior is achieved by selection of organic additives which suppress the plating rate on the surface of the wafer relative to the rate within the feature. Differences in plating rate between the surface and feature base have been attributed to the slower mass transfer rate of current suppressing additive species into the features compared to the mass transfer rate to the surface. More recent filling models are based on the gradual accumulation of current accelerating species within high aspect ratio features as their internal surface area decreases during initial plating [18–21].

Figure 16.6 shows a typical sequence in which a high aspect ratio via with a thin copper seed layer is filled with electroplated copper. Initially the seed layer coverage within the via is only a few hundred angstroms thick, while the thickness on the wafer surface is over 1000 Å. Following electrodeposition for 10 s, the copper growth rate on via sidewalls is approximately equal to the growth rate on the wafer

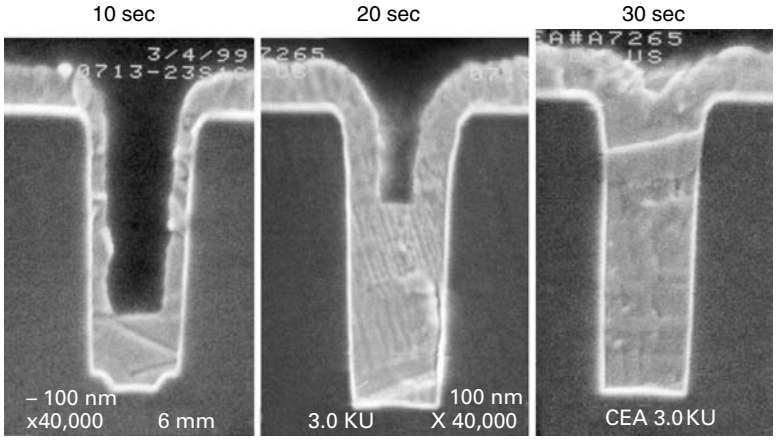


FIGURE 16.6 Deposition of copper in 0.3 μm vias as a function of plating time at a current density of 10 mA/cm^2.

surface and the growth rate at the base of the via has accelerated. Both of these aspects of initial growth behavior are required for successful filling of damascene features by electrodeposition. Initial nucleation and growth on all surfaces is necessary to prevent sidewall void formation, and accelerated growth at the feature base is required to achieve complete fill prior to pinch-off of the feature due to conformal growth at the feature neck. Fill continues from the bottom up, until it is complete after about 30 s. Depending on the organic additives and the current waveform, accelerated growth may continue after filling is complete. This leads to an excessive deposit thickness over superfilled features and increases the difficulty of the subsequent CMP process.

Copper can also be deposited from solutions using electroless and immersion deposition [8,9]. Electroless deposition is analogous to electroplating in many respects, however, the electrons required for cupric ion reduction are supplied by chemical reactions involving a reducing agent such as formaldehyde, thus eliminating the need for power supplies and electrical connection to plated surfaces. Immersion plating takes place when metal ions in solution are reduced to form metal atoms on a surface which itself is oxidized to supply the required electrons, thus limiting this method to very thin film depositions on readily oxidized surfaces. These processes have not been implemented in damascene applications.

16.2 Fundamentals of Electroplating

16.2.1 Kinetics

The relationship equating electromotive force (EMF) or voltage to free energy of an electrochemical charge transfer reaction in solution is given by Equation 16.4:

$$E_0 = -\Delta G/nF \tag{16.4}$$

where E_0, standard EMF of cell reaction ($Cu^{2+} + 2e^- = Cu$) (V); n, number of electrons per atom reacted; F, charge on a mole of electrons (Faraday constant, C/mol); ΔG, free energy of reaction (J/mol).

The units of the terms in this equation are seen to be equivalent since

$$\text{V} = \text{J/C} = \text{(J/Mol)/\{(electrons/atom)(C/Mol)\}}.$$

The standard EMF for an electrochemical reaction such as $Cu^{2+} + 2e^- = Cu$ must be expressed as a voltage relative to some other reaction as explained in detail in electrochemistry textbooks [22]. It has been found that a convenient and reproducible reference voltage (NHE) is created by the reaction of

Reaction	Reduction potential vs NHE
$Au^+ + e^- = Au$	1.68
$O_2 + 4H^+ + 4e^- = 2H_2O$	1.229
$Pt^{2+} + 2e^- = Pt$	1.2
$Ag^+ + e^- = Ag$	0.799
$Cu^{2+} + 2e^- = Cu$	0.345
$2H^+ + 2e^- = H_2$	0
$Pb^{2+} + 2e^- = Pb$	–0.126
$Sn^{2+} + 2e^- = Sn$	–0.134
$Ni^{2+} + 2e^- = Ni$	–0.23
$Co^{2+} + 2e^- = Co$	–0.28
$Cd^{2+} + 2e^- = Cd$	–0.403
$Fe^{2+} + 2e^- = Fe$	–0.409
$2H_2O + 2e^- = H_2 + 2OH^-$	–0.827
$Ti^{2+} + 2e^- = Ti$	–1.63
$Al^{3+} + 3e^- = Al$	–1.76
$Na^+ + e^- = Na$	–2.71

FIGURE 16.7 Electrochemical series of standard E_0 values for reduction of metal ions.

hydrogen gas to form hydrogen ions and electrons on a platinum electrode surface. Based on the definition of the EMF for the hydrogen oxidation reaction as 0.00 V, the standard reduction potentials (E_0) of several metals at 1.0 M solution concentration are given in Figure 16.7. Equilibrium reduction potentials (E) for divalent metal ions shift by about 29 mV per decade of metal ions activity [O] change from the 1.0 M reference point according to the Nernst equation:

$$E = E_0 + (RT/nF)\ln[O] \tag{16.5}$$

where [O], molar concentration of oxidized species.

Metals that are thermodynamically unstable, such as sodium, exhibit negative potentials for reduction of the ion to the metal. The negative potentials correspond to a relatively non-spontaneous positive free energy of reaction. Noble metals, such as gold, which are thermodynamically stable, exhibit positive potentials for reduction of the ion to the metal. The positive potentials correspond to a negative free energy of reaction. As illustrated in Figure 16.8, when a potential more negative than the E_0 value of a metal/ion couple is applied at a surface immersed in a solution containing ions of the metal, reduction of the metal ion to the metal takes place. This is an electroplating process. In the convention normally used for electroplating, the electrons supplied for the metal ion reduction constitute a positive current flow. If the potential is held at the E_0 value no current will flow, while if the potential is held at a value more positive than E_0 the metal will thermodynamically favor dissolution to metal ions resulting in a negative current flow. Numerous factors can influence the rates of reduction and oxidation reactions as the potential of an interface is moved away from E_0. Kinetically fast reactions can allow current to increase sharply with applied voltage changes of less than 50 mV, while slow reactions require larger applied voltages to drive substantial current. Typical electroplating reactions take place at applied interfacial potentials of several tenths of a volt relative to their E_0 value.

Using standard electrochemical conventions, the applied voltages which result in a substantial current flow corresponding to metal ion reduction becomes increasingly negative as the metal ion to be reduced

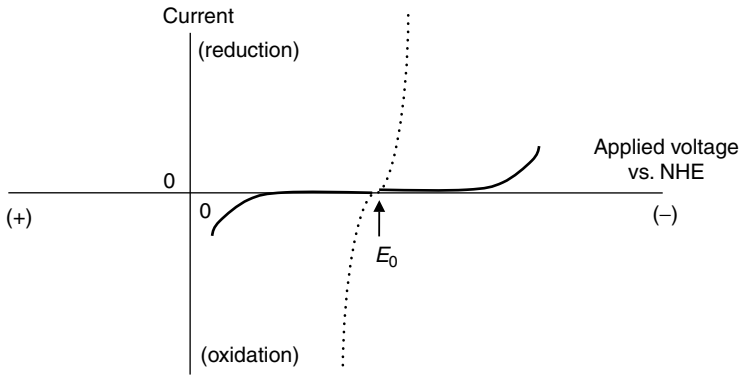


FIGURE 16.8 Current as a function of applied potential for metal ion reduction and oxidation at metal surface with high and low degrees of charge transfer resistance.

becomes less noble (see Figure 16.9). When an electrolyte contains several metal cations which can be electroplated onto a cathode, the voltage which is applied will determine which reactions proceed. As shown in Figure 16.9, a voltage of about −0.3 V relative to Normal hydrogen electrode (NHE) will cause cupric ion reduction. At this applied voltage, free metal ions such as Pt and Ag will also undergo reduction. Ions in solutions such as Ti will not be reduced, and water itself is stable. As a result of this behavior, copper deposition can be carried out in practical systems containing significant levels of metals such as Na, K, Ti, Ca, Fe, and Mg, which are difficult to reduce without resulting in significant deposit contamination. Electrolyte contamination with Ag, Au, or other noble metals will result in their co-deposition when copper plating is carried out.

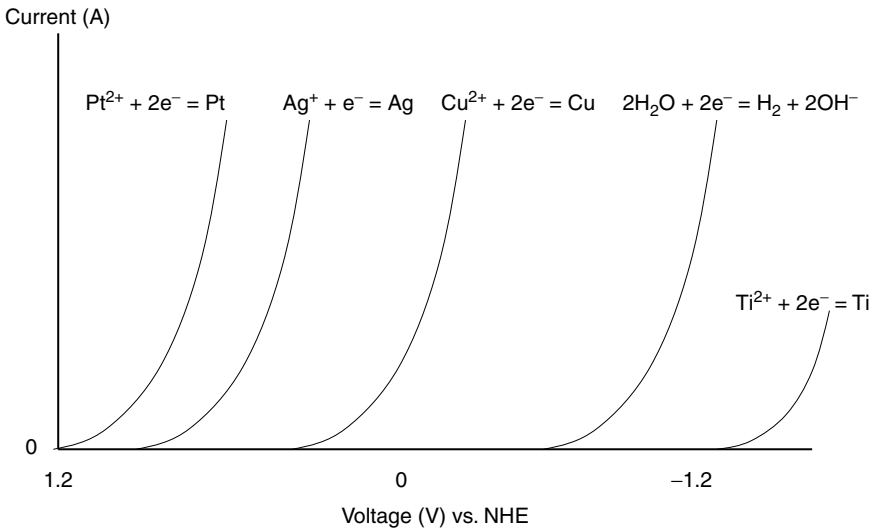


FIGURE 16.9 Current–potential characteristics for reduction of several metal ions and water.

Although plating is normally performed at controlled currents, it is convenient to express current flow at a given point on the cathode surface as a function of applied voltage, interfacial kinetics, and mass transfer using the relationship:

$$I = i_0[O_{(0,t)}]e^{-\alpha nF(E_a - E)/RT} \tag{16.6}$$

where $E_a = \eta + E$, applied potential at metal/solution interface; $O_{(0,t)}$, oxidized species concentration at the interface at a given time and other terms are as defined previously. While this equation is generally difficult to solve, it accurately reflects the effects of applied interfacial potential, concentration of the ion at the interface, and inherent interfacial kinetics on the reduction current at a given point on a surface.

Increases in applied potential result in exponential increases in the deposition current at a given interfacial ion concentration and exchange current. In this respect, the current dependence on potential can be considered analogous to a reaction rate dependence on temperature in an Arhennius equation. Alternatively, the interface can be considered as a resistance element in an equivalent circuit, which decreases in value as the applied voltage increases.

The exchange current determines deposition current sensitivity to applied voltage and interfacial concentration changes. Large exchange currents, which correspond to very rapid and usually reversible metal reduction charge transfer kinetics, result in strong dependence of the deposition current on mass transfer and applied potential. A comparison of exchange current impact on deposition current sensitivity to applied potential is illustrated in Figure 16.8. In copper plating systems, the exchange current can be reduced by deposition at lower temperatures and use of organic additives, which slow kinetics.

The current flow at a given point on the surface is proportional to the concentration of metal ions at the interface regardless of other factors. The interfacial concentration is typically increased by faster mass transfer and higher bulk concentrations of metal ion in solution. Figure 16.10 illustrates metal ion concentration profiles as a function of distance from the cathode at different mass transfer rates assuming a fixed current density. Faster mass transfer corresponds to a higher interfacial ion concentration and a shorter distance from the interface at which the bulk ion concentration is reached. Plots of the cupric ion deposition current density as a function of applied potential are shown in Figure 16.11 for several rotating disk electrode rotation rates. The range of rotation rates includes the range of mass transfer conditions normally employed during wafer processing. At low applied potentials or current, a relatively

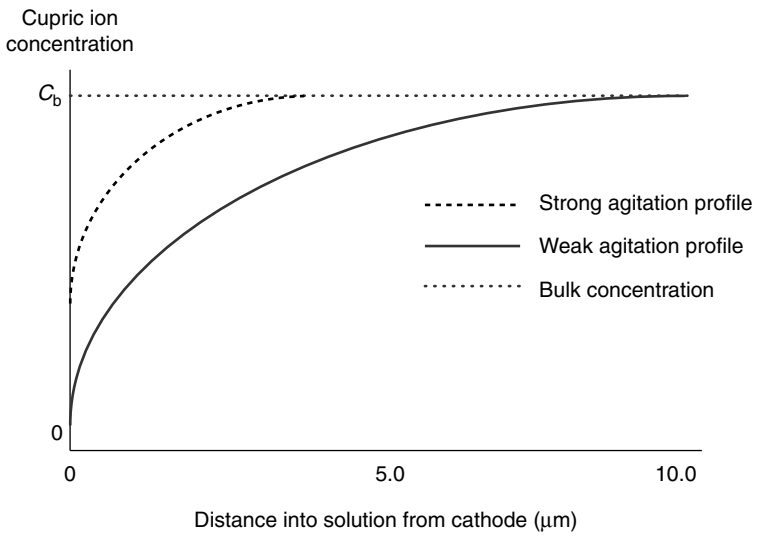


FIGURE 16.10 Cupric concentration profiles in solution at fixed current.

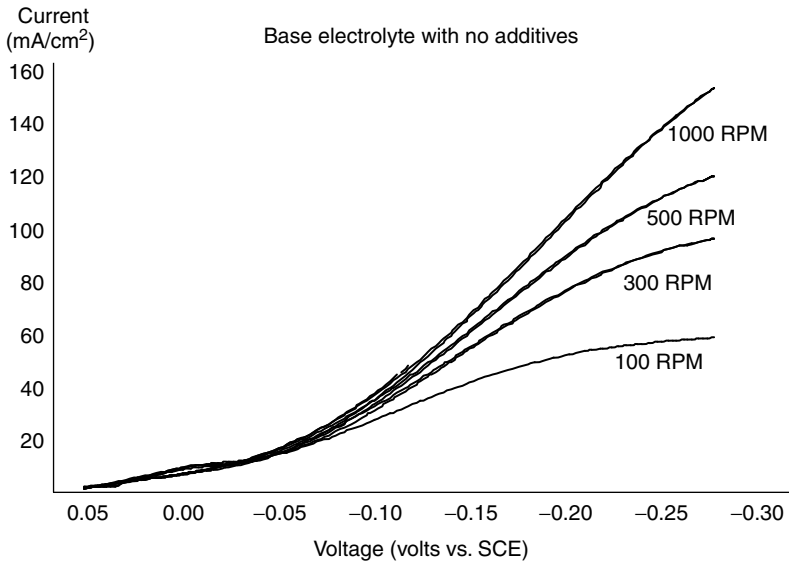


FIGURE 16.11 Effect of mass transfer on I–E relationship.

small dependence of the current density on the mass transfer rate is seen. Gradually, as applied potential or current is increased and larger concentration gradients develop between the interface and the bulk solution, the dependence of the current density on the mass transfer rate increases.

Thus, based on Equation 16.6, uniform deposition rates across large surfaces such as wafers can best be generated using uniform mass transfer and applied potential conditions across the wafer surface. Selection of process conditions with a relatively low exchange current will also enhance thickness uniformity.

16.2.2 Mass Transfer

Mass transfer of metal ions in electroplating solutions takes place as a result of convection, diffusion, and migration mechanisms. Convection includes ion movement as a result of pumping of plating solution to achieve directed flow, mechanical agitation or rotation, ultrasonics, and as a result of thermal gradients. In electroplating systems, convection due to pumping and mechanical agitation is usually primarily responsible for maintaining equilibrium concentrations of species across distances much greater than a distance of about 1 mm. Both laminar or turbulent flow are common, with laminar flow generally desired except when very high rates of mass transfer are required to achieve a high current density and rate of deposition. Increasing convection decreases the thickness of the stagnant diffusion layer at the cathode interface. Typical diffusion layers, across which diffusion and migration of the cupric ions are the primary means of mass transfer, range from a few microns to about 100 μm in thickness.

Migration is the movement of ions which takes place as a result of the potential gradient applied across a solution between the anode and the cathode during plating [22]. As a result of migration, cations will move toward the negatively charged cathode and anions will move toward the positively charged anode. In the bulk of the plating solution, the amount of mass transfer due to migration is negligible. However, near the interface where convection is small, migration of a metal ion can become significant when the charge carrying capacity of the metal ion in solution becomes significant compared to the charge carrying capacity of other ions in solution. This condition takes place when the transport number for the metal

ion, T_i, as given in Equation 16.7, becomes a significant fraction of 1.0.

$$T_i = \frac{C_i Z_i D_i}{\sum C_x Z_x D_x} \tag{16.7}$$

C_i,C_x, concentrations of metal ion and other ionic species respectively; Z_i,Z_x, charge of metal ion and other ionic species respectively, D_i,D_x, diffusion coefficient of metal ion and other ionic species respectively.

In typical copper plating applications, the high concentration and diffusion coefficient of hydrogen ions result in its domination of charge migration behavior. This results in relatively small amounts of migration of cupric ion toward the interface compared to the rate of diffusion. In plating solutions with very low amounts of hydrogen ions, however, mass transfer of cupric ion toward the interface can be increased due to migration.

Diffusion, the movement of ions in solution across a concentration gradient, is relatively fast across distance scales such as the dimension of IC features or the stagnant interfacial diffusion layer. Fur cupric ion, a diffusion coefficient of 5×10^{-6} cm^{-1} [23] results in a 0.1 ms time for diffusion across a 0.1 μm distance, and a 1.0 ms time for diffusion across a 10 μm distance. The short times for diffusion across these distances result in ample replenishment of cupric ion within high aspect ratio features during a typical 5–20 s filling process even though the volume of solution containing sufficient cupric ion to form the copper metal deposit in a feature is 200–400 times the feature volume.

16.2.3 Geometry Effects on Local Kinetics

So far, the deposition process has been considered as if it were taking place at a single point or assuming that all points on a surface are equivalent. Although the basic kinetic and mass transfer behavior can be consistent across a surface, the actual interfacial voltage, which drives the deposition reaction usually varies with position on the surface [24–26]. For example, the interfacial voltage may be different at the center and the edge of a wafer for several reasons. The variation of the interfacial voltage across a plated surface can be broken in to four classes of behavior. First, if part of a plated surface is closer to the anode than another part of the plated surface, then the voltage drop through the solution from the anode to the closer cathode surface is smaller and more voltage is available to drive the deposition reaction at the closer surface. A higher current on the surface closer to the anode results. This behavior is utilized in Hull cell testing to conveniently observe plated deposit appearance across a wide range of current densities, but is not usually a factor in well designed plating cells which maintain a constant anode to cathode distance. Second, the edge of a plated surface will normally plate at a higher rate than an adjacent continuous surface. This behavior is often characterized as a field or antenna effect, however, it is more accurate to consider that the ohmic voltage drop from an anode through solution to an edge will be less than the corresponding voltage drop from the anode to a non-edge surface. Edge effects can be prevented by using a plating cell with internal wall dimensions defined by an insulating cylinder of internal dimension equal to the wafer diameter. Third, when plating only part of a surface to form a defined pattern, there are often variations in the density of plated surface area relative to dielectric covered or non-plated area. In this case, the areas of lower plated density will plate more quickly than dense areas. This effect is similar to edge (antenna) behavior and result as a higher available interfacial voltage is possible due to a reduced voltage drop through the bulk of the solution in the less dense pattern area. Normal pattern density effects are not observed in damascene plating since all surfaces are plated, however, an inverse effect due to the high plated surface area within features is possible in high circuit density areas of the wafer before feature fill is complete. Pattern density and edge effects on thickness distribution can be reduced by using highly conducting electrolytes and low plating currents, both of which reduce voltage drop variability in solution. Fourth, a resistive voltage drop across the anode or the cathode itself can result in higher interfacial voltages and plating rates at surfaces nearest an electrical connection to the power supply. In damascene electroplating this is known as the terminal effect and will be described in greater detail in the section covering damascene plating thickness distribution control.

16.3 Damascene Cu Electroplating Chemistry

16.3.1 Electrolytes

Considerations for choosing an electrolyte include solution conductivity, wetting and oxide dissolution ability, rate of dissolution of the copper seed layer in the solution, adequate cupric ion concentration to support high current density, diffusion and migration behavior of the solution, solubility and activity of additives, cost, and waste treatment capability. To date, electrolytes commonly used for damascene copper applications contain sulfuric acid as a supporting electrolyte, copper sulfate pentahydrate as a source of cupric ions, and chloride ion to modulate cupric ion oxidation and reduction reactions. Among the numerous electrolytes used for non-damascene Cu plating applications [12], damascene related evaluations of as methane or propane sulfonic acid/copper sulfonate, and copper pyrophosphate electrolytes have been reported [18].

The sulfate based electrolytes for damascene applications can be divided into three categories based on their composition and performance as summarized in Table 16.1. High acid electrolytes (175 g/L acid, 17.5 g/L Cu) were initially introduced for all damascene applications based on their high solution conductivity and resulting ability to reduce field or edge effects on thickness distribution in various plating cell geometries. The demonstrated manufacturability of the high acid baths for high-end printed circuit board applications was also considered. High acid solutions were also well suited for use with existing additives and were especially effective in rapidly dissolving copper oxide present on the seed layer prior to plating.

As damascene plating evolved, requirements for higher throughput, reduced seed dissolution, and deposition of uniform films on thinner and more resistive seed layers gradually increased. Low acid and intermediate acid or highly saturated electrolytes were introduced to address these issues. In low acid electrolytes, the concentration of sulfuric acid is reduced to about 10 g/L, resulting in a sharp solution resistance increase. While this is detrimental to antenna or edge effects on thickness distribution, the increase in solution resistance reduces the importance of seed layer resistance and improve can the thickness distribution across the wafer. In addition, when the sulfuric acid concentration is reduced, more cupric ion can be added to the electrolyte without exceeding the solubility product of cupric sulfate. Concentrations of cupric ion in low acid electrolytes that are usually 40 g/L or higher allow deposition at relatively high current density without depletion of cupric ion at the interface. Finally, the low concentration of hydrogen ion in these baths begins to result in migration of cupric ion toward the cathode interface (see Equation 16.7). This helps to maintain a high interfacial cupric ion concentration, although the effect still tends to be small compared to the rate of cupric ion diffusion. Potential concerns with low acid electrolytes include the performance of additives and the rate of copper oxide dissolution upon wafer immersion in the bath. The capability of additives to generate bottom-up fill becomes poor as the acid concentration is decreased below about 5 g/L, probably due to increased ionization and solubility and decreased adsorption of the disulfonic acid accelerator molecule.

A compromise between the low and high acid electrolytes, which has been investigated and occasionally used for damascene plating is often called intermediate acid. Typical compositions contain from 20 to 60 g/L of acid and 28 to 70 g/L of copper. The more saturated solutions approach the solubility of copper sulfate in water. This reduces the dissolution of copper seed compared to high

TABLE 16.1 Electrolytes for Domascene Electroplating

Bath Type	Cupric Ion (g/L)	Sulfuric Acid (g/L)	Chloride (mg/L)	Bath Temperature
High acid	15–20	150–200	25–75	20–30
Low acid	35–70	5–20	25–75	20–30
Intermediate acid	30–70	20–60	25–75	20–30

and low acid electrolytes, and can result in improved nucleation and bottom-void performance on thin copper seeds. Except for the loss of thickness distribution control across highly resistive seeds achieved using lower acid concentration, these electrolytes generally result in good on-wafer performance. Drawbacks of concentrated electrolytes include the formation of excessive amounts of copper sulfate crystals on tool surfaces due to evaporation, and a tendency for anode passivation since the copper sulfate solubility product can be exceeded near the anode surface as cupric ion dissolves and results in a locally high concentration.

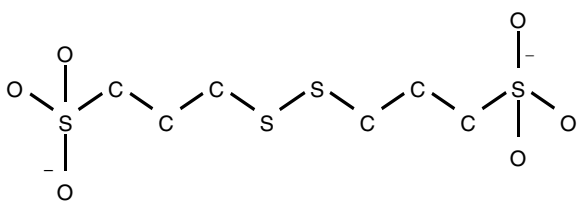
Chloride ions, added as HCl, are present at concentration of 20–100 ppm in all sulfate baths used in damascene applications. Nominal concentrations are usually close to 50 ppm. Although chloride ions are considered to be part of the electrolyte, the primary requirement for their addition to solution is based on interactions with the organic additives required to achieve desired metallurgy and fill performance. By itself, chloride ion adsorbs strongly on the plated Cu surface and provides a charge transfer site which are slightly catalytic or depolarizing in sulfate electrolytes [27]. With the organic polymer suppressor additive present, however, interaction between the adsorbed chloride ion and the suppressor results in formation of a well-adsorbed polymer film which blocks interfacial charge transfer and is therefore highly polarizing [28–31]. Current suppression on the wafer field by this mechanism is essential to achieve bottom-up fill. In addition to the interaction with the suppressor, more recent work [32] has shown that the chloride ion interacts with the accelerator additives in charge transfer reactions during cupric ion reduction. Finally, a uniformly defect-free and smooth plated surface and desired metallurgy are achieved only using chloride ions along with organic additives.

Chloride ions have been reported to serve other functions such as anode depolarization, or to enhance formation of anode films [12], but the importance of these effects is not clear in damascene applications. Bromide ions are known to have effects similar to those of chloride ions in sulfate baths, however, bromide ions are not used with any commercially available additive systems. Iodide ions are insoluble in copper sulfate baths and therefore have little effect.

16.3.2 Organic Additives

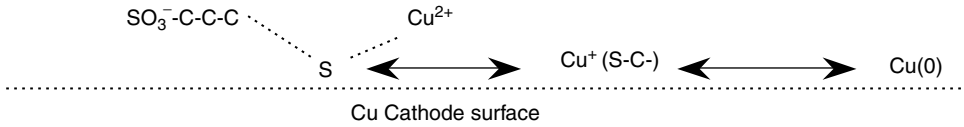
Organic additives are added to copper sulfate plating solutions to achieve the desired metallurgy, defect, thickness distribution, reliability, and fill performance. The terminology used to describe these additives has varied among chemical suppliers and has sometimes been confusing and not related to the specific function or structure of an additive. Here, the additives are referred to as accelerator, suppressor, and leveler species based on the primary function of each of the three additive types.

Accelerators are defined as species which increase the current at a given applied voltage when they are added to a plating bath containing other additives. Accelerators are commonly added to the electrolyte as the sodium or potassium salt of dimercaptopropane sulfonic acid (SPS) at concentrations in the range of 2–20 ppm. The structure of the accelerator molecule, as shown below, results in unique adsorption properties and interactions with the cupric ion reduction process.



In general, sulfide, and thiol like functional groups result in very strong adsorption on copper, gold, and similar metal surfaces [33]. In the case of SPS, the (–C–S–S–C–) portion of the molecule is believed to interact with the copper surface and result in strong adsorption. Normally, strongly adsorbed thiols or sulfides are rapidly incorporated in a plated film, resulting in a short lifetime on the copper surface and high levels of sulfur and carbon in the plated films. With SPS, however, the disulfonic acid functional

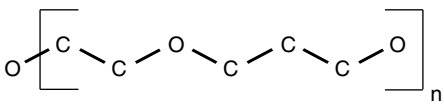
group of the molecule is highly soluble in the plating solution. This characteristic may inhibit incorporation in the plated film, and almost indefinitely extend its lifetime on the plated copper surface. While on the copper surface, SPS and its related monomer, mercaptopropone sulfonic and (MPS), are believed to act as charge transfer sites with a lower energy for the rate limiting step of Cu^{2+} to Cu^{+} reduction [34–38].



Based on the adsorption and charge transfer properties, and the observation that catalytic properties can be imparted by pre-dipping in SPS and MPS containing solutions, these molecules are believed to behave as catalysts participating repeatedly in cupric ion reduction reactions [38]. At the low solution concentrations normally used during plating, the catalytic activity of the molecule can increase with time as accumulation on the plated surface continues [39]. The accumulation of this molecule is also the basis for rapid bottom-up filling and the subsequent overplating as will be discussed in the feature fill section.

The accelerator molecules are subject to a variety of oxidation and reduction reactions at the cathode and the anode [40,41] and can form several by-products and complexes with cupric or cuprous ions. Most of these species are unstable and eventually decompose to form propane sulfonic acid or insoluble molecules.

Suppressors are defined as species which adsorb to reduce the current density at a given applied voltage as they are added to a plating bath containing accelerator additives. Suppressors are normally polyethylene glycol (PEG) or polypropylene glycol (PPG) type polymers or co-polymers with the characteristic structure as shown below.



Molecular weights of suppressors generally range from about 1000 up to 10,000 and solution concentrations generally range from 100 to 1000 ppm, much higher than typical accelerator or leveler concentrations. Because of the relatively high suppressor concentration in solution, suppressor adsorption rapidly results in monolayer-like film formation on the copper surface [31,42,43]. Microbalance [44] and current response studies suggest that film formation is complete in about 0.2 s under typical conditions. The film acts to inhibit deposition current at a given applied potential. Several models for current inhibition by the adsorbed film have been suggested including blocking of growth sites, restriction of cupric ion diffusion to the surface, complexation of reacting ions, and formation of a polymer electrolyte films in which ions are solvated [45]. Figure 16.12 shows current as a function of voltage as potential is cycled from the rest potential to −360 mV vs. saturated calomel electrode (SCE) reference electrode and back. As the suppressor concentration is increased from 100 to 1600 ppm, in an electrolyte containing approximately 2 ppm SPS, the current suppression behavior increases, especially at higher currents and during the scan from −360 mV back to the rest potential. This hysteresis between anodic and cathodic potential scans is characteristic of systems containing accelerator and suppressor additives, which are capable of bottom-up filling.

The degree of suppression achieved depends on the molecular weight of the polymer, its structure, and its relative PEG/PPG content. The strongest current suppression is achieved for polymers with higher molecular weights up to about 10,000 [41], and greater PPG content. Interaction with adsorbed chloride ion greatly increases suppressor adsorption strength. Various Raman and FTIR spectroscopy studies have suggested specific chemical interactions between the C–O bonds in the glycol molecules and the adsorbed chloride [46].

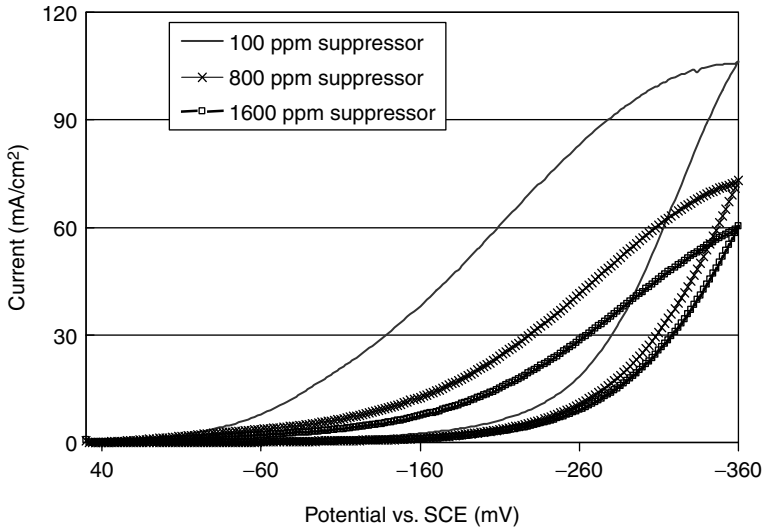


FIGURE 16.12 Polarization curves showing increasing current suppression as polymer suppressor concentration is increases in an electrolyte containing 2 ppm SPS accelerator.

Suppressors also function to reduce the surface tension of the plating solution and improve its ability to wet the copper surface. The surface tension of plating solutions is reduced from about 70 dynes/cm without suppressors added, to 40–60 dynes/cm as suppressor concentration increases. This results in improved wetting behavior as measured by a decrease in the contact angle of the plating solution on the copper seed as shown in Figure 16.13. Low contact angles ensure complete rapid wetting of small features. Assuming reasonably low contact angles, hydrostatic pressure of up to several atmospheres is generated in sub-micron features by capillary action. This results in rapid air dissolution or removal. Suppressor selection for uniform wetting of the copper seed surface is key to avoiding pit defects which can result when air is not fully displaced during initial wafer entry into the plating bath.

Suppressors are reasonably stable in plating solutions in the absence of current flow, but cleave into lower molecular weight fragments as a result of anodic oxidation and other mechanisms during plating.

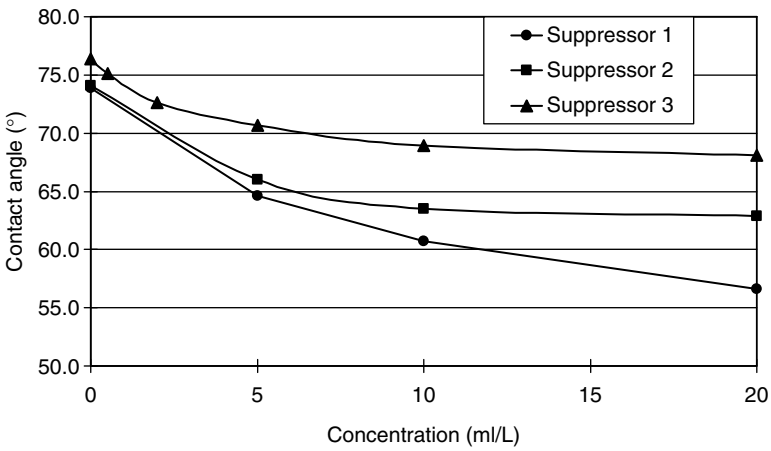


FIGURE 16.13 Contact angle as a function of suppressor concentration for three suppressor molecules.

Usually, the lower molecular weight fragments are less current suppressing than initial suppressor species and become virtually inert at a molecular weight below about 300 [31,41]. Since polypropylene glycol above a molecular weight of about 1200 is not highly soluble in plating baths, it is essential to avoid suppressor structures which can cleave to form this type of fragments.

Levelers are normally known in plating applications as molecules, which adsorb to suppress the rapid plating which may otherwise take place at corners, edges, or other geometric irregularities on a plated surface as a result of field and mass transfer effects [47–49]. Levelers are usually strongly current suppressing nitrogen containing polymers with molecular weights in the range of 1000–15,000. In the acidic plating baths, the nitrogen functional groups may be protonated, leading to a net cationic charge on the molecule. Since the function of levelers is intended to be localized to the rapidly growing surfaces, the concentrations of levelers in solution is kept low, often near or under 1 ppm, and leveler lifetime on the plated surface is ideally low due to either incorporation in the film or decomposition. Under these conditions, leveler surface coverage is diffusion limited, and generally higher at corners, edges, or surface irregularities resulting in slower growth on these surfaces. If the leveler is cationic, migration to highly cathodic edge or corner surfaces may be enhanced, leading to more coverage at sites, which would have a higher current density for normal field effect reasons.

In damascene applications, levelers are added to the plating bath primarily to reduce the rapid copper growth over features which have undergone bottom-up fill, and thus improve the planarity of the film for subsequent CMP processing. In this application, specific chemical interactions in which the leveler neutralizes adsorbed accelerator species present at a high surface concentration over filled features is useful to achieve good deposit planarity. Levelers can also impact fill behavior by either reducing the growth rate of copper at the upper sidewall of a feature to allow better filling, or by adsorbing at the feature's base and slowing otherwise rapid bottom-up filling. Because levelers are often incorporated in the plated film, proper selection is important for film purity, resistivity, and reliability behavior.

16.3.3 Anodes and Anode Films

As was shown in Figure 16.4, a copper metal anode is normally immersed in the plating solution and undergoes oxidation based on the equation $Cu(O) \rightarrow Cu^{2+} + 2e^{-}$. This maintains both chemical and electrical equilibrium in the plating bath. In some copper plating baths, high purity or "oxygen free" copper anodes are used [50–53]. In acidic sulfate baths, however, use of these anodes was found to result in rapid decomposition of organic additives, as well as formation of fine copper particles which contaminated the plating solution and resulted in defect formation in plated films. Use of anodes containing 400–600 ppm of phosphorous was found to solve these problems, and these "phosphorized copper" anodes are most commonly used for damascene applications [53–56]. Except for phosphorus, anode purity is usually 99.99% or higher. The most common impurities are silver, iron, nickel, arsenic, zinc, and tin each of which is typically present at about one ppm or less. Most anode impurities are not readily deposited on the wafer because their reduction potentials are considerably more cathodic than that of cupric ion (see Figure 16.7), although Ag is readily co-deposited with copper.

The reduction in additive consumption and fine copper particulates seen using phosphorized copper anodes is largely achieved by formation of an adherent low density black film which forms on the anodes during plating. The film is believed to restrict additive mass transfer to the anode interface, and inhibit the loss of partially dissolved grains before they are fully oxidized. Black films generally increase in thickness up to several mm as an anode is initially used. In some applications this change can result in variable additive use during tool start-up, and has been addressed by anode conditioning to prepare a black film before tool use. In addition to the phosphorous content of the anodes, suppressor, and possibly other additive components are required for this film formation. Molecular composition of anode films is not well understood, but analysis shows P, C, Cu, S, O, Cl, and sometimes silver as typical constituents of an amorphous structure.

Anode grain structure and the distribution of phosphorus in the anodes can vary among manufacturers using proprietary fabrication methods. Usually, grain sizes are on the order of 25–

100 μm and phosphorous is present at the grain boundaries. This P distribution may also inhibit fine copper grain loss by reducing electrolytic dissolution along the grain boundaries. Phosphorous has also been associated with inhibition of the undesired anodic reaction $Cu(O) \rightarrow Cu^{+} + e^{-}$, although this effect is not well documented.

Damascene plating hardware has been designed to reduce the possibility of fine copper or anode film particles from reaching the wafer [57], however, complete anode particle elimination can be achieved by using inert anodes such as platinum coated titanium. When platinized titanium or similar anodes are substituted for copper in a standard bath, the anode reaction becomes $2H_2O \rightarrow O_2 + 4H^{+} + 4e^{-}$. Difficulties with this approach include the generation of large volumes of oxygen which can form undesired bubbles, rapid oxidation of additives on the platinum anodes, and a decrease of copper concentration in the plating bath during plating. Copper concentration in this type of plating system can be maintained by cupric oxide addition [58,59]. An alternative inert anode approach in which ferric and ferrous iron salts are added to the copper plating bath has also been proposed [60]. In this system, the reaction $Fe^{2+} \rightarrow Fe^{3+}$ takes place at the inert anode to eliminate oxygen evolution, while cupric ion is replenished and Fe^{2+} is replaced in a separated reaction of Fe^{3+} with copper metal. Iron is not co-deposited with copper as a result of its more negative reduction potential [60].

16.4 Damascene Film Deposition

16.4.1 Feature Fill and Leveling Capability and Mechanism

16.4.1.1 Fill Mechanism

Void-free filling is the most basic requirement of copper electroplating for damascene applications. The ability to fill high aspect ratio features using copper electroplating is based on achieving significantly more rapid deposition kinetics near the base of a deep feature compared to the kinetics on the field of the wafer. This allows filling from the base of the feature before conformal plating on the top corners of the features joins together to leave a void. Such capability is usually called bottom-up fill or superfill in the literature, and is achieved by non-equilibrium distribution of current accelerating and suppressing additives between the feature base and the field [13,20,61].

It is useful to consider the current–potential behavior of several plating solutions to understand the capability for bottom-up fill in electroplating. Figure 16.14a shows polarization behavior at a rotating disk electrode of electrolyte containing cupric ions, sulfuric acid, and chloride. Current is measured as potential is first swept negatively from the equilibrium value of +50 mV vs. SCE to −280 mV vs. SCE, and then cycled back to the equilibrium potential. Results are shown for three electrode rotation rates corresponding to different rates of mass transfer. In this simple system, the current increases in an approximately exponential manner until mass transfer begins to limit overall current flow. At 50 rpm limitation of cupric ion mass transfer reduces the overall current magnitude over much of the practical current density range (10–50 mA/cm^2), while at 1000 rpm mass transfer effects on current flow are minimal until current densities above 100 mA/cm^2 are reached. It should be noted that the current at any given potential during the cathodic potential sweep is nearly equal to the current at the same potential during the subsequent sweep toward anodic potentials. This lack of hysteresis between the potential cycles implies that a single current corresponds to each applied voltage regardless of the potential sweep history, and suggests rapidly equilibrating surface morphology and growth characteristics when no organic additives are present. Inability of a plated surface in a given bath to exhibit more than one current at a given applied voltage makes bottom-up filling based on surface kinetic differences impossible in that bath since applied voltages will never be higher near a feature base than on adjacent field surfaces.

Polarization curves using the same electrolyte along with an early generation organic additive formulation, containing accelerator, leveler, and suppressor species are shown in Figure 16.14b. Two major differences are noted in the polarization behavior following additive addition. First, currents are reduced at a given voltage. This behavior reflects changes at the interface caused by the additives including blockage of growth sites, formation of interfacial complexes, topography changes, and film

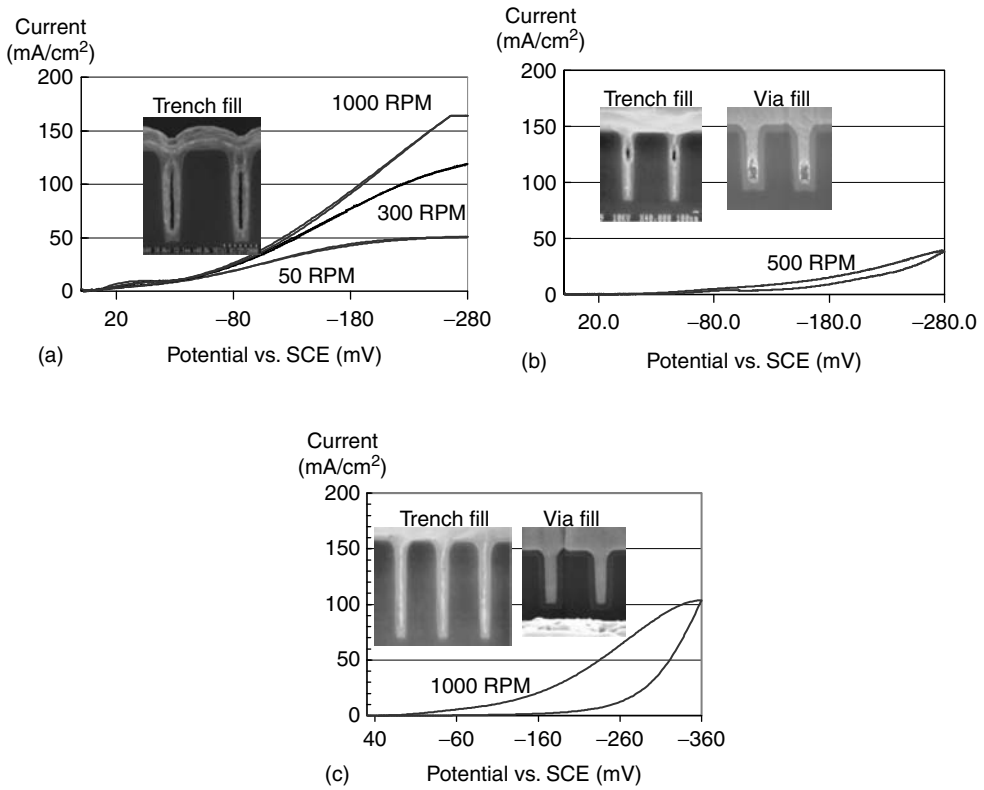


FIGURE 16.14 Polarization and fill results with and without additives. (a) Electrolyte with no additives present. (b) Electrolyte with early damascene additive package. (c) Electrolyte with suppressor and accelerator additives selected for filling performance.

formation, which reduces metal ion mass transfer to the interface. Second, there is a hysteresis behavior in the current flow between the cathodic and anodic potential sweeps such that currents at a given voltage during the initial cathodic sweep are lower than those observed during the subsequent anodic sweep.

In Figure 16.14c, polarization curves measured for a two component (accelerator—suppressor) additive containing bath formulated to achieve very rapid bottom-up fill are shown. This bath exhibits stronger current suppression and more hysteresis between current flows during the cathodic and anodic sweeps. The degree of hysteresis reflects the ability of a plating bath to support both high and low currents at a given applied voltage, and thus can suggest capability of a bath to provide bottom-up fill [19,61–65]. During the initial cathodic potential sweep, the currents measured using this chemistry are also significantly lower than those shown in Figure 16.14a and b. This polarization behavior difference results in higher voltages being applied in order to achieve a given initial current flow. Higher applied voltages are well known to be more effective in achieving uniform nucleation on variable surfaces and to prevent corrosion of metal. As a result, these more polarizing electrolytes are considered useful in achieving growth on discontinuous seed layers. Highly suppressed currents on the wafer field can allow increased growth rate differentiation between the field and the feature base as long as the feature base is not also strongly polarized.

Trench fill and via fill results obtained by plating in the additive free, early generation, and bottom-up fill baths are also shown in Figure 16.14a–c. The trench fill results provide a good means to evaluate the bottom-up fill acceleration capability of a plating chemistry on a well seeded surface. Without additives present, the copper deposition rate on all surfaces within the feature is approximately equal. This leads to

conformal growth and a large center void extending much of the length of the 5:1 aspect ratio trench. This result is observed over a current density range of approximately 5–50 mA/cm^2. At current densities above approximately 50 mA/cm^2, depletion of cupric ion within the feature further limits growth within the feature and the void volume increases further. Fill performance in vias with discontinuous or very thin and oxidized PVD copper seed provides a good means to evaluate the nucleation capability of a plating bath. In the relatively non-polarized bath with no additives, a large void at the base of a via (bottom void) as shown in Figure 16.14a is observed. This reflects an initial dissolution rate or Ostwald ripening of the thinner seed, which exceeded the initial plating rate.

When early generation additives are added to the electrolyte, complete filling of 5:1 AR trenches is achieved, thus demonstrating the existence of bottom-up fill. A center void is, however, observed following plating of 6:1 or higher aspect ratio, 0.17 μm trenches as shown in Figure 16.14b. The bottom void performance is improved somewhat on a discontinuous seed as compared to results for the bath with no additives, but a void is still observed if the seed coverage is poor.

Using strongly current suppressing additives, which result in increased current–voltage hysteresis, full fill of 9:1 AR trenches is achieved, as shown in Figure 16.14c. This rapid bottom-up fill capability has been incorporated in all commercially available additive systems now in use for IC copper circuitization. Via fill is also complete, indicating the additional applied voltages required to drive a given current early in the process which were sufficient to improve nucleation and growth on the thinner areas of the seed to a rate which exceeded seed dissolution. While plating in highly suppressed plating baths is useful to add margin to the filling processes, it is generally necessary to optimize seed processes to yield continuous films to guarantee consistent void free filling and optimum reliability.

Based on the hysteresis behavior in mixed additive systems, and the current accelerating and suppressing characteristics of the additives alone, it is possible to generate bottom-up fill based on two mechanisms. According to the first mechanism, which can be descriptively named "field suppression", the additives quickly adsorb on the field of the wafer to for a current suppressing film. Within the features, however, current suppressing molecules fail to form a suppressing film due to diffusion limitations or steric hindrance based on similarity of the suppressor and feature sizes [4,66–69]. Current–potential characteristics leading to fill based on this mechanism are illustrated in Figure 16.15. Fill performance based on this mechanism can be estimated based on electrochemical studies comparing I–E behavior on an electrode using a complete additive system to simulate deposition on the field, with I–E behavior without additives present to simulate deposition at the feature base. A fill acceleration ratio

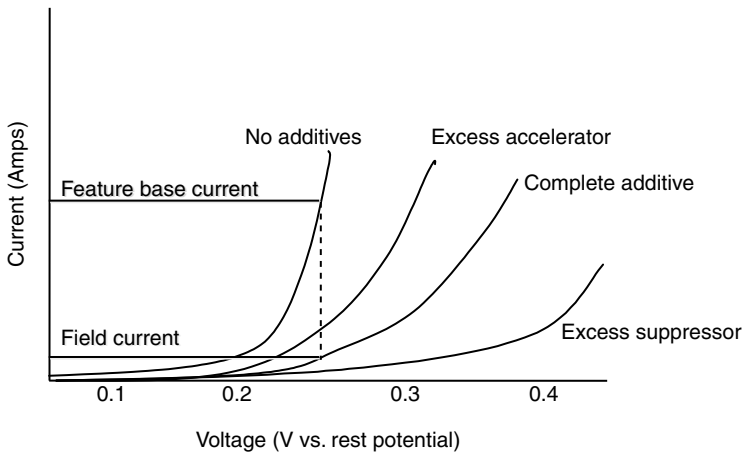


FIGURE 16.15 Polarization characteristics of wafer field and feature base which lead to bottom-up fill based on lack of suppressor diffusion into feature.

TABLE 16.2 Polymer Size and Diffusion Coefficient vs Molecular Weight

Polymer MW	Spherical Conformation Diameter (nm)	Diffusion Coefficient (cm^2/s)
100	0.62	0.7×10^{-5}
1,000	1.32	0.25×10^{-5}
10,000	2.85	0.11×10^{-5}
1,00,000	6.2	0.05×10^{-5}
10,00,000	13.2	0.025×10^{-5}

at a given applied voltage can be calculated by comparing the current for the no additive case to the current for the case with all additives present.Ratios as high as 30:1 are often calculated based on comparing rotating disk electrode data for complete additive systems with the data without additives present. In a practical case, it is unlikely that suppression by additives at the feature base will be zero since diffusion distances and times are on the micron per millisecond scale and polymer additives, normally of molecular weight under 10,000, are smaller than typical feature sizes as shown in Table 16.2. A variation of this mechanism involves limitation of chloride, rather than polymer, mass transfer into small features. Since adsorbed Cl is required for strong polymer based current suppression using typical PEG suppressors, high currents within features could result from lack of adsorbed chloride [70].

According to the second mechanism, which can be descriptively named "accelerator accumulation", additives initially adsorb with reasonable uniformity on the copper seed surfaces both on the field and within the features on the wafer [19,20]. Deposition begins conformally on all surfaces according to this mechanism. As conformal growth continues and features are partially filled, the surface area within small features decreases, especially at the feature base corners. This decrease in surface area is believed to concentrate the strongly adsorbed accelerator additives, leading to displacement of the polymer suppressor and increased current within the features. Current–potential behavior leading to this fill mechanism is illustrated in Figure 16.16. Like Figure 16.15, the field I–E behavior can be approximated by measuring the full additive system using a rotating disk electrode. In this case, however, the current near the feature base is estimated by measuring current–potential behavior at an electrode immersed in an electrolyte containing accelerator only. Current acceleration ratios of about 10:1 are typically obtained

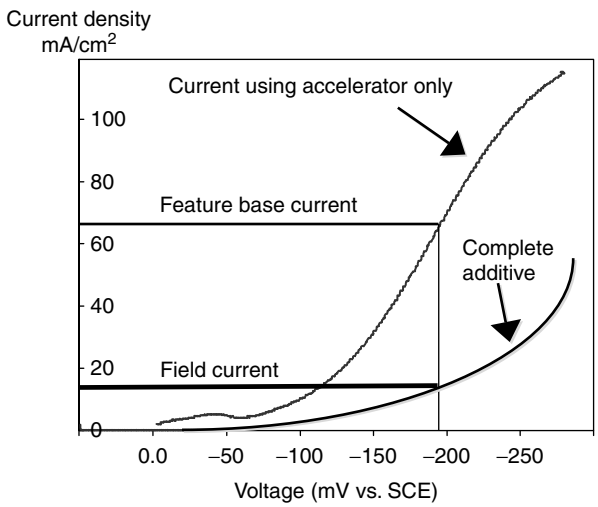


FIGURE 16.16 Polarization characteristics of wafer field and feature base which lead to bottom-up fill based on a depolarization of suppression within features due to accelerator accumulation.

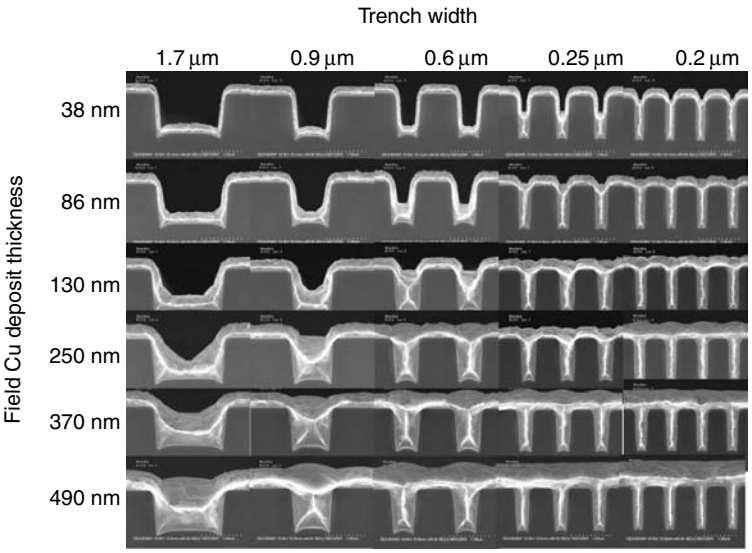


FIGURE 16.17 Fill evolution in trenches as a function of trench width and copper thickness deposited on the wafer field.

over a range of applied voltages comparing the simulated field current to the simulated feature base current. This is in good agreement with measured fill acceleration rates obtained by comparing growth rate within a feature adjacent field growth rates. Measurement of hysteresis during potential cycling at electrodes can also be used to approximate current density or applied voltage ranges over which optimum bottom up fill can be generated [64].

16.4.2 Fill Evolution Behavior

Figure 16.17 shows fill evolution in 1.2 μm deep trenches of 0.2, 0.25, 0.6, 0.9, and 1.7 μm widths following electrodeposition of 38, 86, 130, 250, 370, and 490 nm of copper on the wafer field at a nominal current density of approximately 10 mA/cm^2.

After electrodeposition of 38 nm of Cu, the total Cu thickness on the field is approximately 188 nm when the thickness of the 150 nm seed layer is considered. At this point in the filling sequence, the largest features (0.6–1.7 μm) show conformal growth with a 80 nm layer of copper on all surfaces within the trenches. In the 0.25 μm trenches, the sidewalls in the upper portion of the trench show conformal growth, however, a distinct surface of copper growing parallel to the base of the trench has formed about 0.6 μm above the base of the trench. This surface is characteristic of accelerated bottom-up growth. Once formed, this surface moves rapidly up the feature, ultimately yielding complete fill. This filling process has nearly been completed in the 0.2 μm trenches.

After electrodeposition of 86 nm, fill acceleration in the corners of 0.9 and 1.7 μm trenches has become evident. The 0.6 μm trenches now show a flat growth surface which has progressed to about 0.5 μm from the trench base. Approximately 600 nm of vertical fill has been achieved in the 0.25 μm trenches during the 48 nm of deposition which takes place between the 38 and 86 nm nominal field deposition. This is a 12.5× acceleration relative to the field thickness increase of 48 nm, and suggests a current density of approximately 125 mA/cm^2 on the localized bottom-up growth surface. The 0.20 μm trench size shows growth which has proceeded above the field by approximately 100 nm, forming the characteristic overplating bump which can follow bottom-up filling.

Between 86 and 130 nm nominal deposition, the vertical growth in the 0.6 μm trench is approximately 600 nm. This is a 13× acceleration relative to the nominal growth rate, similar to that achieved in the 0.25 μm features. After 130 nm deposition, more pronounced corner growth is also noted in the 0.9 and 1.7 μm trenches, and fill of the 0.6 μm trench is complete.

After 250 nm deposition, further corner growth is noted in the 1.7 μm trenches, and the 0.9 μm trenches have undergone the transition from accelerated corner growth to rapid bottom-up fill to form the characteristic flat growth surface. Longer deposition times result in similar transitions to rapid bottom-up growth fronts in features as large as approximately 3.0 μm in a 1 μm deep dielectric. The exact delay time, or nominal deposit thickness, required for the transition to accelerated bottom-up fill varies with nominal current density for all feature sizes. Very low current densities never result in transition to bottom-up fill, and currents above an optimum value generally increase the nominal thickness required for the transition.

The feature size dependence of initiation of bottom-up fill acceleration and the observation of initially accelerated corner growth are consistent with the requirement of a threshold surface area change within the feature prior to fill acceleration. During the initial conformal deposition, localized concentration of the strongly adsorbing accelerating species takes place. Accumulation of accelerator is initially most pronounced in the corners at the base of trenches where surface area change is also initially significant. In features between 0.3 and 0.9 μm in width, the total surface area within the trenches has decreased by 40%–50% at the point when rapid growth acceleration begins. The feature size dependence of the timing of bottom-up filling is difficult to explain solely in terms of a filling mechanism based on diffusion limited suppression.

16.4.3 Fill Response to Plating Chemistry

Although the electrolyte and additive formulations used for IC filling applications vary, several trends in the response of fill to additive chemistry have been commonly observed [19]. In general, all filling baths contain cupric ion, sulfuric acid, chloride ion, accelerator, and suppressor. Leveler is not required for fill, but may be present.

Strongly accelerated fill has not been achieved in electrolytes containing suppressing polymers alone or with chloride ion when no accelerator is present in solution. Growth in these baths appears to be largely conformal except at very low Cl ion concentrations where some evidence of filling has been reported [70]. On the other hand, bottom-up fill is most pronounced for systems, which typically contain 50–200 ppm suppressor concentrations along with suitable accelerator additives and chloride ion concentrations. It is clear that suppressors are required to reduce the current on the field of the wafer in order to achieve bottom-up fill. Optimum filling is observed when the concentration, diffusion behavior, and adsorption strength of the suppressor are selected so that suppression on the field is maximized, while suppressor adsorption within features is not strong enough to prevent localized accelerated growth. Very strong suppressors and high concentrations of most suppressors will prevent bottom-up filling by uniformly suppressing growth both within the features and on the field, however, certain suppressors can yield good fill performance at concentrations as high as 2000 ppm.

As the chloride ion concentration is increased from zero, fill acceleration increases from near zero to a maximum value and then decreases as the chloride concentration is further increased. Typically, optimum fill is observed when Cl concentrations are in the range of 30–50 ppm and fill becomes more conformal as Cl concentrations approach 100 ppm. Chloride concentrations sufficient to enhance polymer suppressor adsorption on the field are required for fill. Excessive Cl can diminish fill, possibly by resulting in excessive suppression within features. The maximum fill rate is usually achieved when the Cl concentration is reduced to a point where polymer suppression remains adequate on the field, but becomes limited at the feature base due to a Cl concentration gradient within the feature.

As the accelerator concentration is increased from zero to the 5–25 ppm range, bottom-up fill increases from near zero to a maximum rate at typical suppressor concentrations. Fill becomes conformal as accelerator concentration is further increased. No fill is observed using an accelerator if a suppressor is not present. An accelerator additive is required to achieve the rapid growth within features which takes place while the field remains suppressed. Too much accelerator, however, results in a loss of growth rate difference between the field and the feature due to a lack of field suppression. The recognition that accelerator additives which preferentially adsorb (relative to suppressors) on copper surfaces could be concentrated within small growing features yielding faster local deposition kinetics was key in explaining the ability to achieve very rapid bottom-up filling.

For a given additive system, too low a copper concentration diminishes the bottom-up fill capability in high aspect ratio features simply due to depletion of cupric ion within the features. Successful filling applications have used as little as 17.5 g/L cupric ion. More commonly, concentrations in the range of 40–50 g/L are employed for optimal fill performance.

Sulfuric acid is not known to participate in filling, however, fill performance does vary with acid concentration. At low sulfuric acid concentrations, the mercaptosulfonic acid accelerator molecule becomes ionized in solution and fails to adsorb strongly on the copper surface. As a result, higher accelerator concentrations are required for bottom-up fill when the acid concentration is decreased to the 5–30 g/L range and little filling can be achieved above a pH of about 2.0. The acid concentration may also result in changes in fill capability due to changes in the suppressor adsorption behavior, however, no general trend has been reported.

16.4.4 Leveling after Fill

Accelerated growth in electrolytes containing a two component (accelerator/suppressor) additive system continues over a damascene feature following bottom-up fill. This growth can result in over a micron of excess copper thickness above features which underwent bottom-up fill, compared to the adjacent field copper thickness. This growth is most pronounced over large arrays of closely spaced lines, which act to cumulatively concentrate accelerator species on the wafer surface following fill. Figure 16.18 illustrates the growth behavior of copper during and after bottom-up fill with and without leveler added to the plating bath. Ideally, the addition of a leveler component serves to suppress current on the rapidly growing fill surface once it protrudes above the field, without impacting the growth rate of the copper

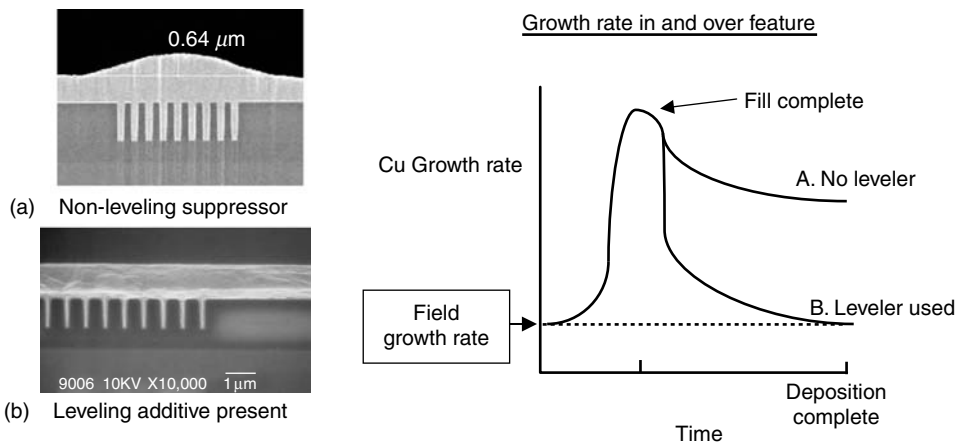


FIGURE 16.18 Copper growth rate in and above a feature during bottom-up fill with and without a leveler added to the plating solution.

within the feature. In practice, the addition of excessive leveler often results in a loss of bottom-up fill, probably due to diffusion of leveler into the filling feature. Development of levelers with diffusion and adsorption characteristics which reduce impact to growth within features remains an area of strong interest. Using modern levelers in the 2–20 ppm concentration range, the copper thickness over dense features following fill is typically less than 1000 Å thicker than the thickness of the adjacent field. The mechanism of overplating reduction by a leveler has not been well studied. However, a specific interaction of the leveler to neutralize adsorbed accelerating species is suggested by its activity at very low solution concentrations and by the relatively small impact of same levelers on the polarization behavior of additive containing solutions.

16.4.5 Fill Behavior in Alternate Electrolytes

Fill performance has been evaluated in non-sulfate based electrolytes including propanesulfonic acid–copper propanesulfonate [18] using additives which yield bottom-up fill in sulfate electrolytes. Little change from the behavior in sulfate electrolytes was noted either in the optimized bottom-up fill rate or the ability to deposit on thin or discontinuous seed. The similarity in via bottom void formation observed using organic and inorganic anion based electrolytes, despite the expected wetting differences, suggests that the wetting limitations of the electrolyte do not contribute strongly to poor electrodeposition behavior on a discontinuous seed.

Rapid bottom-up fill has not been reported using alkaline electrolytes such as pyrophosphate or cyanide in which the cupric ion must be strongly complexed. In these electrolytes the deposition kinetics are limited by complex dissociation, which makes differences in kinetics due to additive adsorption less significant.

16.4.6 Current Waveform and Mass Transfer Impact on Filling

Damascene features are usually filled using direct current (DC) controlled processes in which the field current density is in the range of 3–15 mA/cm^2 (0.07–0.36 μm/min). Lower currents have been observed to result in both bottom voids due to seed dissolution and poor bottom-up acceleration, possibly because the times for filling are so long that additive adsorption on the field and within the feature equilibrates to the same levels. A higher field current density than 15 mA/cm^2 during filling necessitates an extremely high current density within the features in order to maintain a high fill efficiency. This results in copper depletion in the feature, and can cause fill to take place over a time scale shorter than required for desired additive adsorption.

At the beginning of the process, current or voltage can be applied to the wafer immediately as it enters the plating solution, or after several seconds of immersion in the plating solution. To prevent dissolution of the seed, in production processes wafers usually enter the plating bath with a current or voltage applied. However, bottom-up fill rate on sufficiently thick seeds is usually not strongly effected by long immersion times before plating begins.

Pulse reverse plating has been applied successfully to improve fill performance of certain additive systems [69,71]. Such plating methods are believed to accelerate dissolution of copper at the opening of high aspect ratio features or to cause the adsorption behavior of the additives to vary in a way that enhances filling. Fill performance using pulse plating has been found to respond in generally the same way as DC plating to process variables and additive concentration changes, suggesting that the basic fill mechanisms remain the same.

The rate of mass transfer of electrolyte to the wafer surface can influence the rate at which additives adsorb. Fill performance has generally been found to vary only slightly within the flow and wafer rotation conditions used for damascene applications. This suggests that the surface behavior of the accelerator and suppressor additives is generally more important than their rate of mass transfer to the surface. Mass transfer rates can be modulated to control leveler adsorption during plating, and higher mass transfer in leveler containing baths usually reduces the amount of overplating.

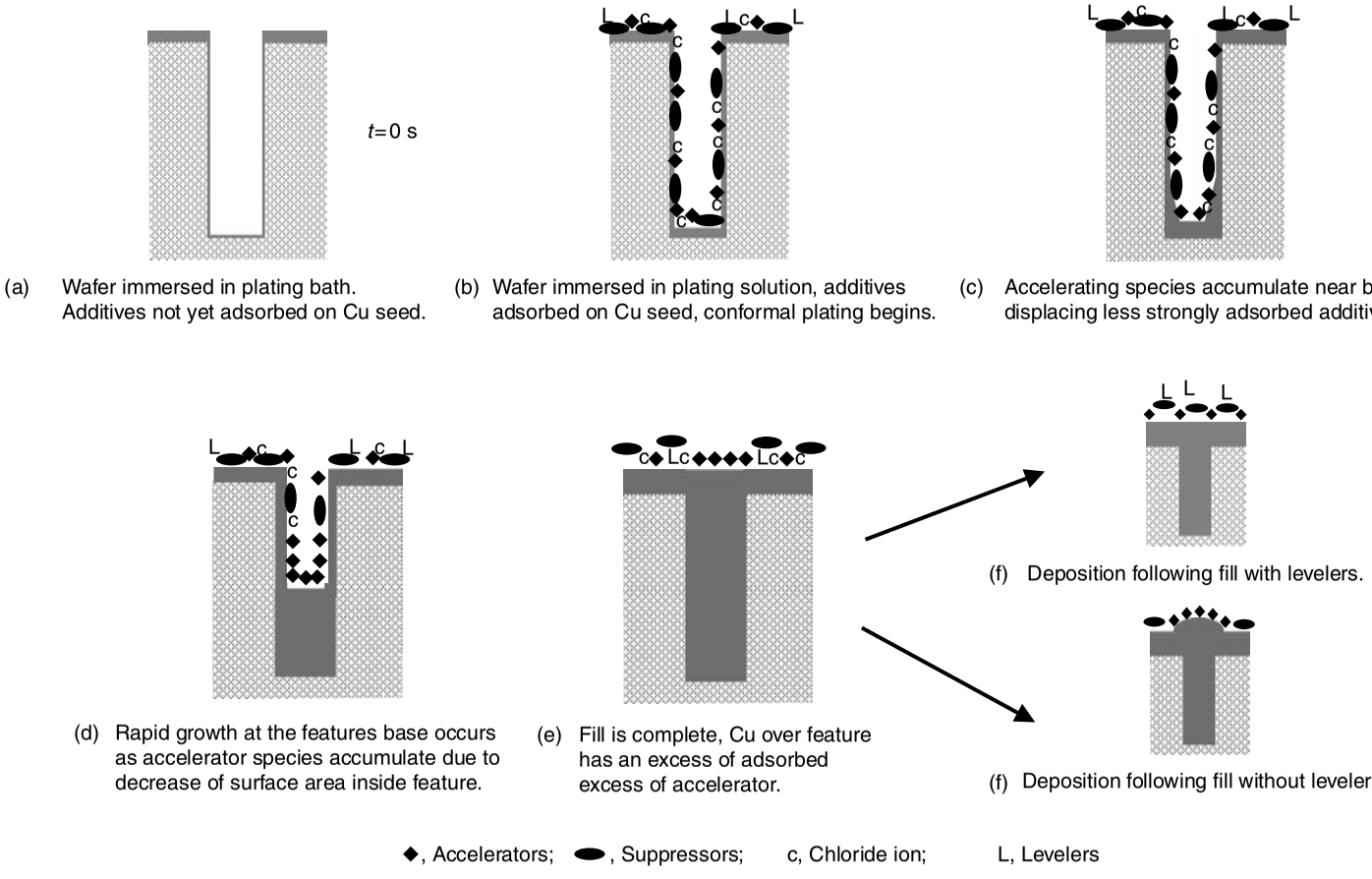

t=0 s
(a) Wafer immersed in plating bath. Additives not yet adsorbed on Cu seed.
(b) Wafer immersed in plating solution, additives adsorbed on Cu seed, conformal plating begins.
(c) Accelerating species accumulate near base displacing less strongly adsorbed additives.
(d) Rapid growth at the features base occurs as accelerator species accumulate due to decrease of surface area inside feature.
(e) Fill is complete, Cu over feature has an excess of adsorbed excess of accelerator.
(f) Deposition following fill with levelers.
(f) Deposition following fill without levelers.
◆, Accelerators; ⬬, Suppressors; c, Chloride ion; L, Levelers

FIGURE 16.19 Fill evolution during bottom-up fill based on accelerator accumulation at the feature base. ◆, Accelerators; Suppressors; c, Chloride ion; L, Levelers.

16.4.7 Summary of Additive Behavior during Filling

Additive adsorption behavior during bottom-up filling can be described based on the dependence of fill evolution on feature size, fill rate and electrochemical suppression behavior as a function of additive concentrations, and current and mass transfer effects on filling. This additive adsorption behavior during filling is illustrated in Figure 16.19a–f.

When a wafer is first immersed in the plating solution, a concentration gradient of suppressing or accelerating species may exist between a feature base and the field. This can happen when the quantity of additive species required to form an adsorbed layer on the surface within the feature, exceeds the amount of additive contained in the solution volume within the feature, or when the suppressing species are large compared to the feature width. This effect could account for less suppression at the via base by slowly diffusing polymer species when relatively low suppressing polymer concentrations (<100 mg/L) are present in solution. Although diffusion limitations at the feature base effect may enhance fill in some cases, the lack of a strong fill rate dependence on wafer immersion time prior to application of current, as well as the relatively long times of plating required for bottom-up fill to begin in large features suggest that this is not critical. Assuming this, by the time significant current flow begins, the adsorption of chloride ion, suppressor, and accelerator species will take place across all field and within-feature surfaces as shown in Figure 16.19b. This should result in similar currents on all surfaces.

After a period of time which depends on current density, feature size, and chemical concentrations, the initial conformal growth leads to a decrease in surface area within features, especially at the bottom corners. This phenomena causes accumulation of the accelerator additives or their accelerating by-products and displacement of suppressor, and results in the initially fast growth observed in corners at the start of the filling process as shown in Figure 16.19c. At this corner surface, the local current density is believed to undergo transition from the lower current to the higher current observed during a cyclic potential scan exhibiting strong hysteresis. Excessively strongly adsorbing suppressor molecules, which prevent bottom-up fill, apparently cannot be displaced by accumulating accelerator and can inhibit this transition.

Once bottom-up fill begins, a growth front moves rapidly upward within features and continues to accumulate accelerating additives as shown in Figure 16.19d. The growth rate within the feature during this stage of filling reaches a maximum and continues until the feature is filled as shown in Figure 16.19e. In the absence of levelers, accelerators remain concentrated over features even after filling and rapid growth diminishes slowly so that overplating above features becomes significant as shown in Figure 16.19f. Addition of levelers at proper concentrations neutralizes the effect of the accelerator as fill within the feature approaches planarity with the field so that overplating is largely eliminated.

16.4.8 Deposit Planarity and Surface Roughness

While overplating above dense features can increase topography variation across a die, the overall planarity after copper plating is also strongly influenced by dielectric etch depth as illustrated in Figure 16.20. Although the copper electroplating process is effective in filling features with aspect ratio's

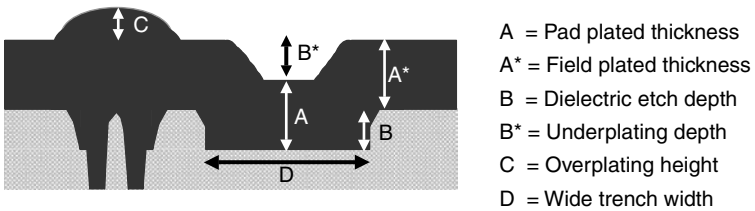


FIGURE 16.20 Overall topography after copper plating of damascene circuitry patterns with super-filled small features as well as large conformally plated pads.

as low as about 1:5 at an accelerated rate, large pads plate at a deposition rate nearly equal to the field deposition rate. As a result, the amount of copper deposited for a given metal layer must at least equal the dielectric etch depth, and the Z position of the copper over large pads after plating is below that of the field by approximately the dielectric etch depth. As a result of this behavior, and the need for a planar copper topography after CMP, plating thickness in the range of 1.3–2.0 times the dielectric etch depth of a given metal level is typical in production.

Plating processes, which deposit little copper on the field while fully filling large pads have been described [72], and continue to be of interest as a result of the associated reduction in CMP cost. These processes involve a combination of electrodeposition and mechanical contact with the wafer surface during electrodeposition rather than simple process or plating chemistry changes. To date, such processes have not been widely implemented.

Plated copper used for IC application usually has a bright reflective appearance with a reflectivity of at least 130% that of bare Si using 460 nm light. The surface roughness of the copper, as measured by Atomic force microscopy (AFM) is usually in the range of 4–8 nm root mean square (RMS). Roughness variability has not been associated with yield or electrical properties of the copper, but is often controlled in order to maintain a reproducible ability to monitor defects using visible light based tools. As feature dimensions decrease into the 45 nm range it becomes increasingly important that roughness evolution remains small compared to the internal dimensions of lines in order to avoid intermittent voids corresponding to rough deposition protruding from the upper sidewall of features during bottom-up filling.

16.4.9 Thickness Distribution

The goal for most damascene copper deposition processes is to form a film of uniform thickness across the wafer. Specifications in the range of 1.5%–2% one sigma thickness distribution, determined from at least 49 point resistivity or laser-acoustic measurements on blanket wafers, are usually met or exceeded in production. Thickness measurement on damascene pattern wafers is complicated by the varying contribution of copper within features to the overall thickness measured using typical methods. In some applications, plated copper is specified to be edge thick or edge thin to match a specific CMP capability. In order to control the thickness distribution, damascene plating equipment has been optimized to control the terminal effect, field, and flow influences on deposition kinetics at a given location on the wafer.

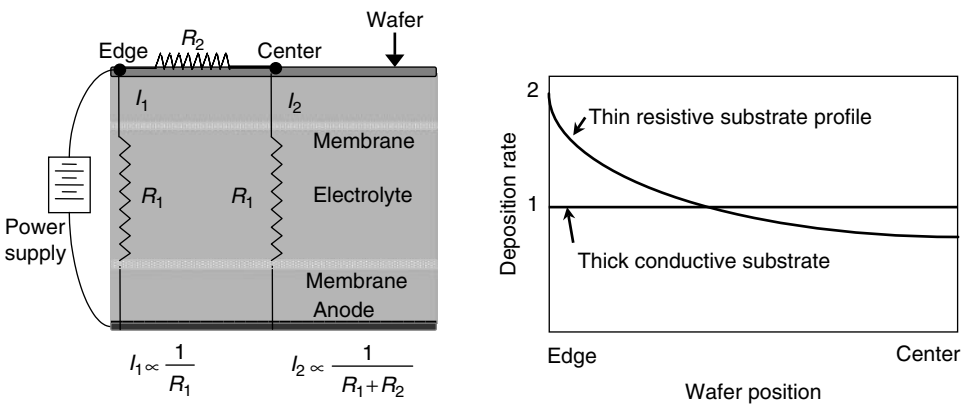


FIGURE 16.21 Circuit diagram of plating cell illustrating terminal effect impact on current density between wafer center and edge.

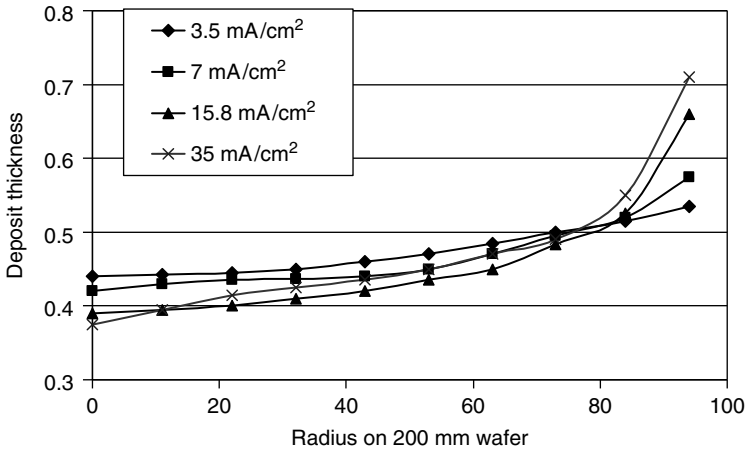


FIGURE 16.22 Plated thickness as a function of radial position for deposition on 100 Å thick (4 Ω/sq) copper seed as a function of plating current density.

16.4.10 Terminal Effect

When copper deposition was introduced in the mid-1990s seed layer thicknesses were generally in the range of 1500–2000 Å, the wafer size was 200 mm, and plating was not usually carried out at high current densities. Under these conditions, resistive voltage drops across the seed layer were too small to influence plating rates. As illustrated in Figure 16.21, when seed layers become thin, the resistive voltage drop through the seed layer from contact points at the edge of the wafer to the wafer center can become significant. This results in a higher interfacial potential at the wafer edge. From polarization curve data (see Figure 16.11), it is seen that interfacial potential differences between the wafer center and edge as small as 10–20 mV begin to impact local current density and deposits will become center thin. Thickness distributions showing the impact of the resistance drop across the wafer while plating on 100 Å thick seed at several current densities are shown in Figure 16.22. A somewhat lower impact of this terminal effect is observed at lower currents since the voltage drops across the seed layers are less under these conditions.

In a plating system, the overall resistance between all points on the wafer and the anode surface should be uniform in order to achieve uniform current densities. Based on this, a number of hardware and process approaches have been used to diminish the terminal effect in semiconductor processing equipment.

First, increasing the resistivity of the plating solution by lowering the acid content has a pronounced impact on the thickness distribution on resistive seeds [17] as discussed earlier. Typical high acid bath conductivity of 0.5 $(\Omega\text{-cm})^{-1}$ results in a cell impedance of about 0.1 Ω and overall cell voltages on the order of 1–4 V using standard plating cell geometries. When bath conductivity is decreased to the range of 0.05 $(\Omega\text{-cm})^{-1}$ by lowering the acid concentration, the relative change in potential between the wafer center and wafer edge compared to the overall potential drop through the plating cell becomes small and the thickness distribution is improved. The plating bath can be considered to behave like a large swamping resistor in an electrical circuit.

A second method, relying on the principle of increased cell resistance involves placing continuous highly resistive membranes or insulating surfaces in the plating solution adjacent to the wafer. Often, an effective resistivity of 100× the bath resistance can be achieved by using suitable materials. This highly resistive layer dominates overall plating cell impedance behavior and effectively forces an equal current density through all portions of the layer. If the layer is sufficiently close to the wafer there is little possibility for current to redistribute between the wafer center and edge and the thickness distribution

across the wafer is improved. The drawback of this method is the 30–60 V which may be required to drive current through a resistive layer, several times higher than the voltages normally used during plating.

A third method of modification of thickness distribution is the use segmented or multiple anodes [73]. Using this method, one anode is normally positioned below the center portion of the wafer and one or more additional electrically isolated anodes are positioned concentrically around the central anode. Current applied to the anodes is then controlled individually. To create a uniform current density on a thin seed, a higher voltage is applied to the anode nearer the wafer center compared to other anodes. Because there is a lower resistance through solution between the center anode and the wafer center compared to the wafer edge, current will flow preferentially to the wafer center. The drawback of this method is the complexity of power supply and anode cell design and maintenance.

A fourth method of across wafer thickness distribution control is the placement of insulating shielding in the plating bath near the wafer edge to restrict current flow to this portion of the wafer. This method can be highly effective in generating uniform films when seed thickness and plating thickness are constant. However, a shield configuration, which yields uniform current density across the wafer on thin seed, will result in too much deposition near the wafer center on partially plated films or thicker seed. As a solution to this problem, the use of dynamic and asymmetric shields has been suggested [74]. Such methods have the drawback of additional moving parts near the wafer and increased mechanical complexity.

A fifth method involves additional cathodes or "thieves", which are placed near the wafer edge. These surfaces, which can be controlled either by the same power supply as the wafer or with an additional power supply, help divert current from the wafer edge. Generally, this method is effective over short distances compared to the wafer diameter and is more effective in controlling distribution near the wafer edge than across the full wafer.

Finally, it was noted earlier [14] that plating bath compositions with very slow deposition kinetics effectively introduce a large and uniform electrical resistance at the wafer interface. This resistance acts to redistribute current as described for resistive membranes placed near the wafer. To achieve interfacial resistance increases, sufficient to counter the terminal effect, the copper concentration in sulfate baths must be sharply lowered or electrolyte compositions, in which the cupric ions are complexed, must be used. To date, these bath compositions have not been shown to have sufficient superfill performance for use early in the plating process.

16.4.11 Field Effects

An ideal plating cell configuration, which avoids field effects during copper deposition on wafers, is a cylinder, which tightly surrounds an anode and a wafer of equal size which are placed at opposite ends of the cylinder. In this configuration, all field lines will be parallel and of equal density across the plating solution and uniform deposition will result in the absence of terminal effects or flow differences. In actual semiconductor equipment it is necessary to have some open space in a plating cell beyond that enclosed by a wafer diameter cylinder. This results in additional field lines or current pathways through the solution and a tendency for a higher current density near the wafer edge. Plating equipment is generally kept as close to an ideal cylinder design as possible. Insulating surfaces (shields) are often placed in the solution around the wafer perimeter to correct for edge effects resulting from non-ideal cell design by redistributing current toward the wafer center.

16.4.12 Mass Transfer

Most wafer plating systems achieve relatively uniform solution flow across the wafer surface using a combination of wafer rotation and pumped electrolyte flow directed toward the wafer at a uniform rate through a manifold or membrane. To the degree to which uniform flow cannot be achieved, mass transfer rates of cupric ion and additive species to the wafer surface vary with wafer position, and plating rates can exhibit corresponding variability (see Equation 16.6). The direction and degree of rate variation,

TABLE 16.3 Properties of Electroplated Copper

Property	Typical Value
Elongation to failure	10–25%
Stress	±50 MPa
Hardness	110–130 Knoop at 1 g load; 1.5–2.8 GPa
Tensile strength	50 KPSI
Modules of Elasticity	100–150 GPa
Reflectivity (vs. Si)	130% @ 460 nm
Roughness	4–11 nm RMS
Grain Size	~1 μm for 1 micron film
Resistivity	1.75–1.8 μm-Ω-cm
Purity	99.98–99.998
Texture	20–70% <111>

however, depends on the plating chemistry and the nominal current density. In the simplest case, too little mass transfer to an area of the wafer results in depletion of cupric ions, and a reduction of the plating rate. This effect becomes more pronounced at higher nominal current densities, in baths with lower cupric ion concentration, and when using lower nominal flow rates. In some cases, depletion of cupric ion near the wafer edge can be used to balance field effects, which tend to produce thicker films at the edge of the wafer [75]. Differences in the mass transfer rate of additives to different location on the wafer surface can impact the thickness distribution independent of cupric ion depletion. Baths with high concentrations of leveling additives can exhibit a reduced current density in areas of high mass transfer. In leveler-free baths the degree of accelerator adsorption on the wafer surface can the impact plating rate; in this case an increased current density in areas of higher mass transfer is observed. Like the terminal effect, the impact of mass transfer differences on the thickness distribution is always proportional to the overall impedance differences between each point on the wafer and the anode. As a result, resistive electrolytes and resistive elements placed near the wafer will generally reduce the impact of mass transfer differences on thickness.

16.4.13 Metallurgy and Reliability

For optimum reliability and electrical performance copper films with large grains, high density, and low or controlled levels of impurities are preferred. Electroplated copper meets these requirements, however, the desired metallurgy is a complex result of deposition current, additive selection, and film annealing conditions. Properties of plated copper are summarized in Table 16.3 and Table 16.4 for films annealed at room temperature.

Ductility, hardness, and tensile strength are rarely monitored in copper deposited for IC applications. Stress in electroplated Cu films can vary from slightly compressive to slightly tensile, and normally approaches zero during room temperature anneal. Relatively high stresses of 200–300 MPa can be generated in plated Cu by annealing the copper above the plastic deformation temperature of about 230°C. Purity of plated copper varies depending on plating chemistry, mass transfer, and current conditions. In baths without a leveler, or in baths containing a leveler which is not incorporated in the

TABLE 16.4 Electroplated Copper Impurity Levels

	Approximate Copper Impurity Concentrations (rpm)					
Additive Type	C	S	Cl	O	N	All Metals
Standard 3 component additive	15–50	5–20	30–100	1–10	1–10	<1.0
Additive system in which no leveler is co-deposited during plating	1–5	1–5	1–5	<2	<2	<1.0

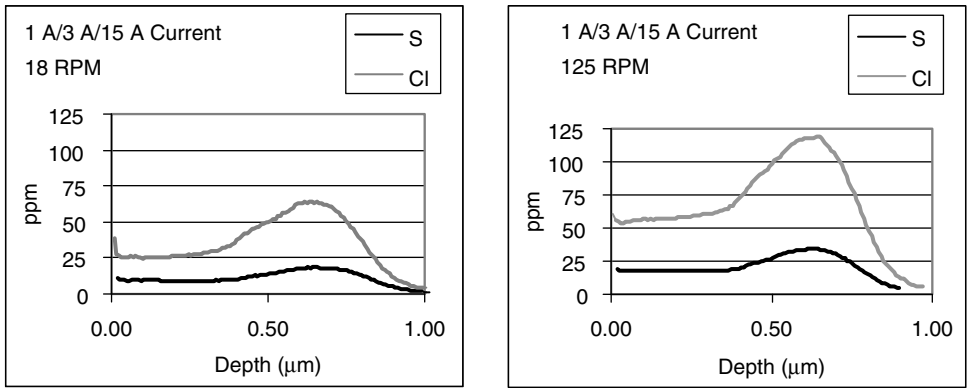


FIGURE 16.23 Copper film purity as a function of depth for films plated at an initial current of 1 A for 11 s followed by a current of 3 A for 30 s and a current of 15 A for 38 s. Films plated using 18 and 125 rpm wafer rotation rates are shown.

plated film, very high Cu purities are achieved. Most levelers used in damascene applications co-deposit during plating to yield significant levels of Cl, S, and C in the copper films. The degree of incorporation always increases with higher mass transfer rates and usually increases at lower plating current densities, both conditions which allow more leveler to diffuse to the wafer surface per unit of copper deposited. This behavior often results in purity variations as a function of depth in the plated films as shown in Figure 16.23. In this example, the upper portion of the film is plated at a higher current density, and exhibits much lower impurity levels than the lower portion of the film plated at a low current density. Higher impurity levels are also observed in the film plated using the higher wafer rotation rate of 125 rpm compared to 18 rpm as the result of greater mass transfer of leveler to the wafer surface. Purity measurement of copper within small features, which undergo bottom-up filling has been difficult to perform accurately because the analysis spot size exceeds the feature dimension for high accuracy measurement methods. The high localized currents during filling and somewhat more limited mass transfer of leveler into small features is expected to lead to an in-feature purity similar to that of copper plated at high currents. Texture of plated copper is more complex than that of Al since <111>

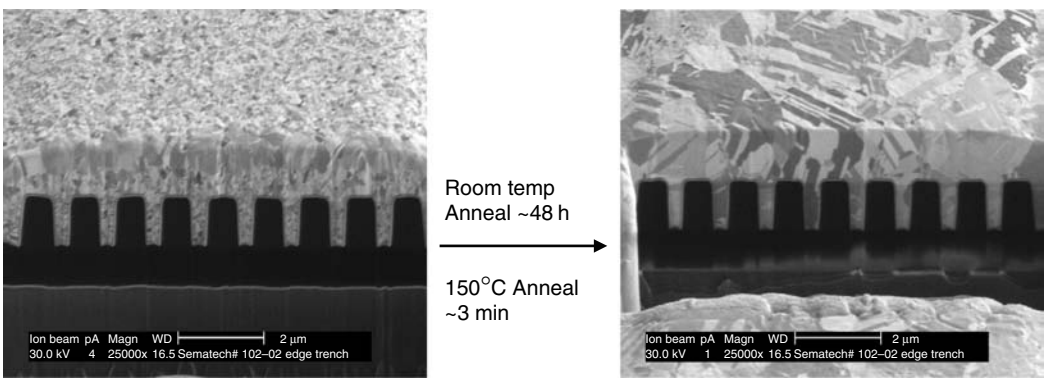

Room temp
Anneal ~48 h
150°C Anneal
~3 min
Ion beam pA Magn WD 2 μm
30.0 kV 4 25000x 16.5 Sematech# 102–02 edge trench
Ion beam pA Magn WD 2 μm
30.0 kV 1 25000x 16.5 Sematech# 102–02 edge trench

FIGURE 16.24 Focused ion beam (FIB) images of grain structure of plated copper films immediately after deposition and following anneal at room temperature.

orientation can be with respect to the field or the feature sidewall [76]. For copper, development of a large grain size, rather than a specific orientation, may be more relevant to reliability.

16.4.13.1 Grain Size and Anneal Behavior

Without additives present, or with a suppressor additive alone, copper generally grows to form columnar grains which follow the seed grain structure. With accelerator and suppressor additives present, the nucleation characteristics during copper growth are changed dramatically such that grains are continually formed during deposition. This results in a very small nominal grain size. The actual grain size after plating has been difficult to verify because of the unstable nature of the grains. Results such as those shown in Figure 16.24a for an ion beam cut of a plated film obtained within 15 min of deposition suggest as plated grains have dimensions of approximately 0.05–0.10 μm. Following deposition, electroplated copper films undergo grain growth even at room temperature to form grains which have dimensions as large or larger than the thickness of the plated film [77,78]. Figure 16.24b shows grains after room temperature anneal for the film shown in Figure 16.24a. In this example grains have grown to approximately 1.0 μm dimensions indicating about 99.9% of the initially plated grains have been consumed by other grains. Films as thick as 80 μm have been shown to undergo analogous growth behavior. This behavior is noted with films over a very wide range of impurity contents [79]. The stress, ductility [77], hardness [80], and resistivity [78] properties of the plated copper shift along with the grain growth, with fully annealed films being more ductile, softer, and about 20% more conductive.

Generally, the grain growth rate increases sharply as temperature of annealing is increased 1 μm thick films achieve nearly complete grain growth after as little as 60 s at 250°C. An activation energy of 19 Kcal/mol for grain growth of copper has been calculated based on stress, ductility, and resistivity change [77]. Anneal rates have been found to depend strongly on plated film thickness, with thinner films requiring a much longer time to undergo complete anneal, and films under 0.3 μm in thickness sometimes failing to undergo anneal at room temperature almost indefinitely. The time to anneal plated copper films has also been found to decrease as copper seed thickness decreases for a fixed plated thickness. In most cases films of higher purity undergo more rapid anneal at room temperature than lower purity films, possibly as the result of less grain boundary pinning by the impurities [79]. Plots of resistivity change as a function of time during grain growth at room temperature are shown for one micron films with high medium and low purity levels in Figure 16.25. After sufficient anneal, the resistivity difference between high and low purity films is generally found to be less than 2% and is

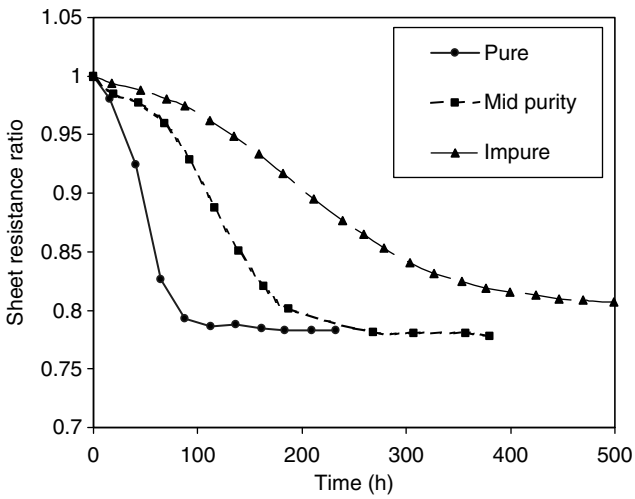


FIGURE 16.25 Resistivity changes during room temperature anneal of plated copper of several purity levels.

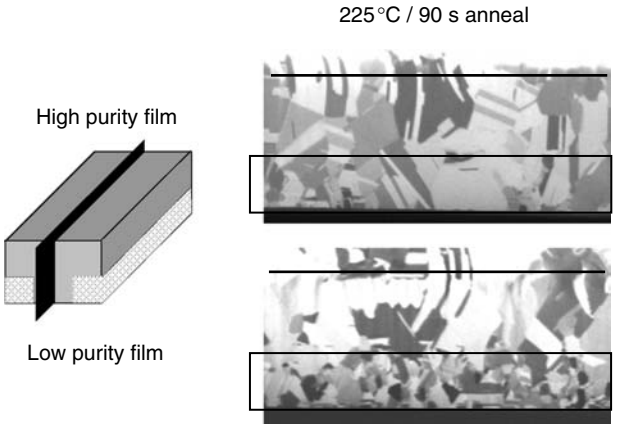
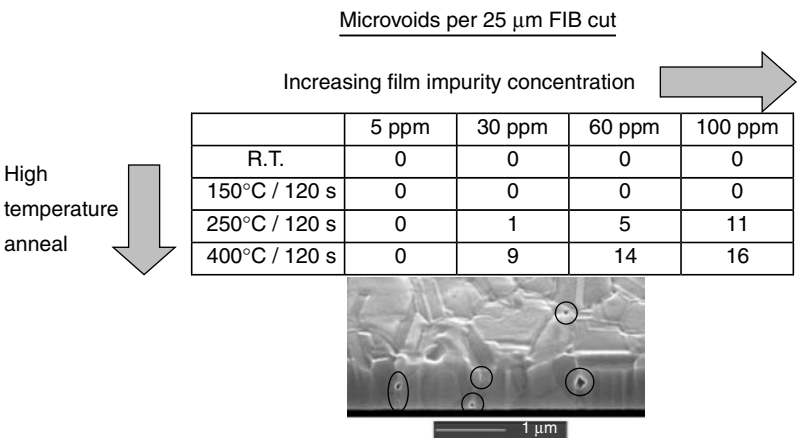


FIGURE 16.26 Cross section along length of 0.2 μm wide, 1.0 μm deep line showing growth of large grains in fine lines for higher purity films, but small grains remaining in lines for lower purity films.

difficult to measure. Anneal rates have also been found to increase for copper films plated at higher current density, independent of their purity [78]. In these cases, the as deposited films have also been noted to show higher initial resistivity possibly suggesting smaller as deposited grains with lower thermodynamic stability.

The grain growth, which electroplated copper on the wafer field undergoes, has been found to drive grain growth within adjacent small features as shown in Figure 16.26, potentially providing improved resistivity and reliability performance. As a result of this behavior, it is usually advantageous to anneal plated copper films prior to CMP of copper to obtain optimum EM and resistivity performance. Nearly all manufacturing process flows follow this anneal sequence and an anneal capability is frequently included in copper electroplating tools. Low purity films may require a higher temperature and a longer anneal time to achieve grain growth within features.



	5 ppm	30 ppm	60 ppm	100 ppm
R.T.	0	0	0	0
150°C / 120 s	0	0	0	0
250°C / 120 s	0	1	5	11
400°C / 120 s	0	9	14	16

FIGURE 16.27 Microvoid formation following anneal of copper films of high and low purity.

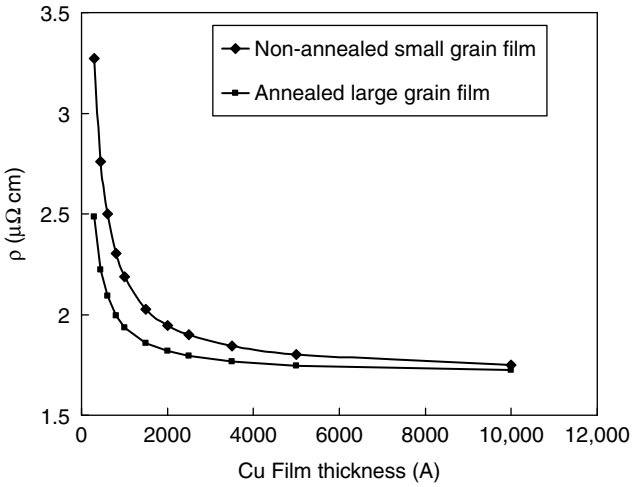


FIGURE 16.28 Copper resistivity as a function of film thickness for highly annealed large grained and small grained films.

During anneal and grain growth it has been found that small voids, frequently called "microvoids", begin to appear in electroplated copper films [81]. Microvoids are usually in the range of 0.02–0.05 μm in diameter and appear at grain boundaries or triple points. Examples of microvoids in a continuous copper films are shown in Figure 16.27 as a function of the anneal temperature and film purity. Microvoid density and size is directly related to the plated copper film purity. High purity films develop only occasional microvoids following 400°C anneal while low purity films develop microvoids with increasing size and frequency as the anneal temperature is increased above 200°C. Chloride impurity levels have been suggested as being most strongly correlated with the tendency to form microvoids, however, the nature of the relationship between impurities and microvoids is not clearly understood. Microvoids can appear within features, however, their formation appears to require consolidation of numerous grain boundaries making formation in large continuous copper volumes more likely.

16.4.13.2 Resistivity

The transition from aluminum to copper interconnects was driven in part by the relatively low copper resistivity of 1.8–2.0 μm-Ω-cm which could be achieved in 90–130 nm generation features. As features become smaller and exhibit physical or grain dimensions similar to the 35 nm electron mean free path in copper, resistivity begins to increases. Figure 16.28 shows copper resistivity as a function of the planar film thickness for films with grains of dimension similar to the film thickness, and for films with uniformly large micron-size grains. The thin large grained films were specially prepared by CMP of fully annealed 1.0 μm thick films. The resistivity of the copper increases sharply as film thickness decreases, for both the very large and small grained films at film thickness below about 30 nm. This behavior, primarily due to electron surface scattering, suggests similar behavior will be seen for Cu resistivity in sub 30 nm features regardless of grain size. In the film thickness range between 50 and 100 nm the large grained films show a considerably reduced resistivity, suggesting grain boundary scattering can be important in features of similar dimension. Similar behavior has been reported and modeled as a function of line dimension [82,83]. The overall importance of fine line resistivity remains unresolved since fine lines may tend to be of relatively short length making their impedance dominated by capacitive, rather than resistive, effects. From a plating process perspective, annealing condition, field thickness, and purity of the plated copper will be important in extending the low resistivity nature of copper to smaller feature

dimensions, although the physical limit imposed by surface scattering is already beginning to impact 45–65 nm generation resistivity characteristics.

16.4.13.3 Stress Migration and Electromigration

Electromigration behavior in copper interconnects depends strongly on the barrier/seed and dielectric cap properties [82], however, the purity of electroplated copper and the grain size within features can also influence EM life. Electromigration life for high purity films is usually about three times longer than that of low purity films [84] for the ranges of impurity shown in Table 16.4 above. Work has suggested that the microvoids present in low purity films migrate to, and move along interfaces, during EM testing eventually accumulating and leading to earlier failure of the lower purity films [84]. Anneal conditions sufficient to grow large copper grains, but at a low enough a temperature so as not to induce significant microvoid formation, have also been found to improve electromigration life [85].

Efforts to improve EM of copper by alloying it with metals have been reported [86,87], however, the degree of improvements have been similar to the impact of the normal film impurities. Most alloying elements cannot be added to copper at concentrations over about 1% without impacting line resistance. This limits somewhat the use of bulk film alloys as a means to improve EM.

Like EM, stress migration life of interconnects depends most strongly on barrier/seed adhesion to dielectric and overall integration schemes. A failure during stress migration testing is usually an open due to missing copper near the base of a via located on a relatively large pad. Stress migration test failures are thought to be the result of stress induced vacancy movement toward or atomic migration away from the high stress location near the via base [88,89], often combined with interface failure between the barrier/seed or dielectric and the copper line or pad. Higher copper purity [90,91] has been found to be associated with earlier stress migration failures. Stress measurements of copper during thermal cycling has shown that high purity films have somewhat greater strength or lower ductility and accumulate more stress than the low purity films [90], probably explaining the relationship of stress migration life to film purity. It has also been suggested that impurities in the plated film act to reduce the diffusion of vacancies under stress, thus reducing the void formation and failure [91]. Based on the high microvoid content of low purity films which show generally improved stress migration behavior, it is unlikely that microvoids migrating from the bulk film are the primary cause for the stress migration failures.

When performing stress migration tests on high aspect ratio features, voids related to incomplete filling can be present in features prior to stress migration testing. In this situation, stress migration testing can be a sensitive measurement of fill quality because the fill related voids will randomly migrate during testing to the via base or sidewall leading to failure.

16.4.13.4 Defects

Defects related to copper electroplating can be classified as pitting or missing metal, abnormal copper growth, and particles. Most serious are pitting or missing metal defects which are the result of plated copper failing to deposit on the copper seed layer [92,93]. These defects can lead to opens in the damascene circuitry following CMP. Several examples of missing metal defects in plated copper films are shown in Figure 16.29. Causes for pit-type defects include incomplete initial wetting of the copper surface, air bubbles adhering to the plated surface, contamination of the seed with suppressor-additive-related polymer agglomerates, foams, micelles, or other particulates that deposit on the seed during entry into the plating bath, pre-existing contamination of the seed, and highly variable of severe oxidation of the copper seed which limits wetting. The size of pit-type defects can range form approximately 0.1 μm, to several microns in diameter. Defects related to air bubbles, poor wetting, or failure to displace air from the copper seed are usually circular and frequently appear in lines corresponding to wetting fronts or air bubble movement across the wafer surface. Pitting defects are reduced by controlling the exposure time of the wafers to ambient air between seeding and plating [92], selection of non-agglomerating suppressor additives with good wetting and low surface tension behavior [94,95], and selection of methods for wafer entry into the plating solution which create a single splash-free wetting interface across the wafer [94].

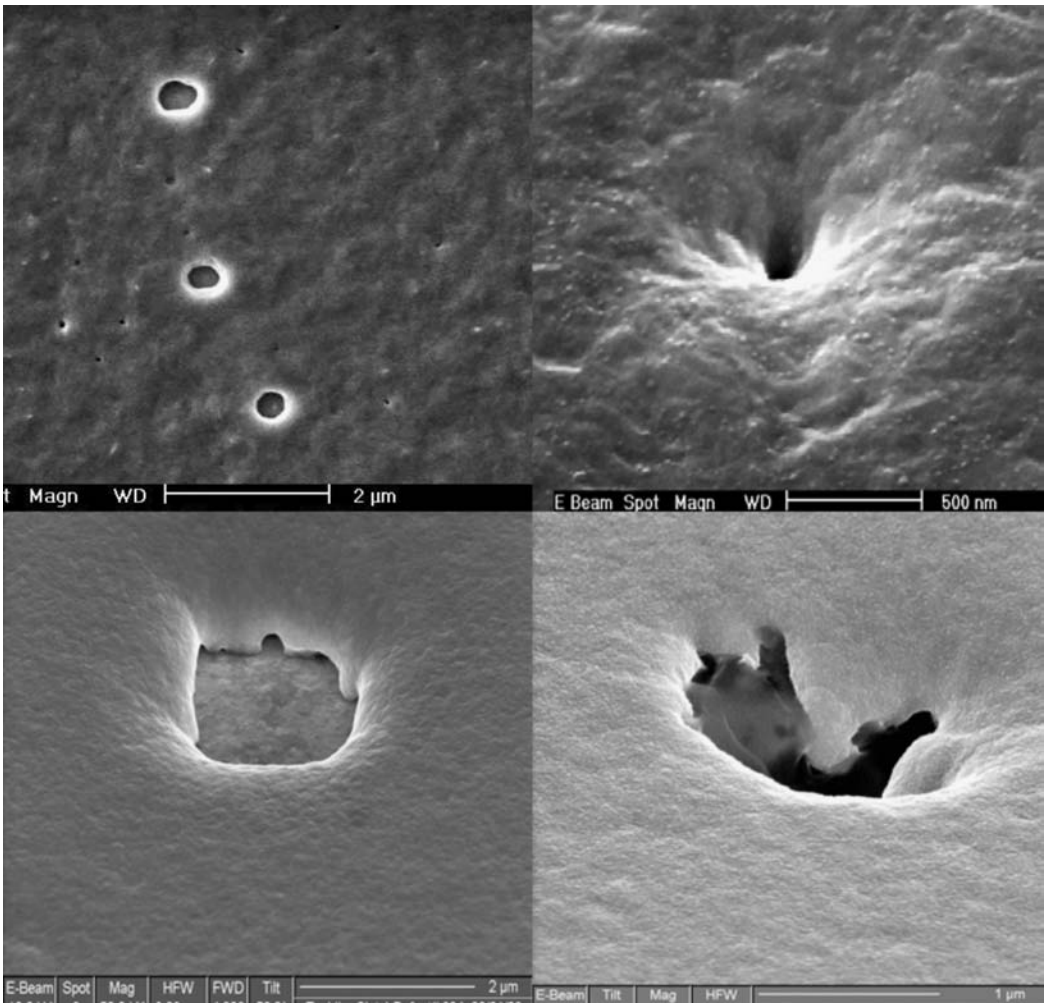


FIGURE 16.29 Missing metal and pitting defects observed after copper plating.

Missing metal defects which become apparent only after CMP can result from formation of microvoids during post-plate copper anneal or from galvanic corrosion during CMP processing.

Abnormal copper growth defects can be detected using various optical light inspection methods following copper plating [5]. In most cases, these defects will be removed during CMP processing and not result in a yield impact. Defect inspection methods do not normally differentiate between pitting and abnormal growth defects, so beyond the desire to have an imperfection-free film, these defects need to be eliminated to facilitate post plating wafer inspection to detect pitting. Abnormal growth defects are usually the result of either non-homogeneous additive adsorption or adherence of small conductive particles to the wafer surface during plating. These particles act as nucleation sites for excessive growth. Defects related to non-homogeneous additive adsorption are usually addressed by the proper selection of additive types, plating currents, and mass transfer rates. Defects related to small particles are most commonly the result of nanometer scale fragments of copper dislodging from the plating cell contact hardware or from the copper anode, and their subsequent transfer to the wafer surface. To prevent anode related particles from reaching the wafer surface, low porosity or ion selective membranes to prevent

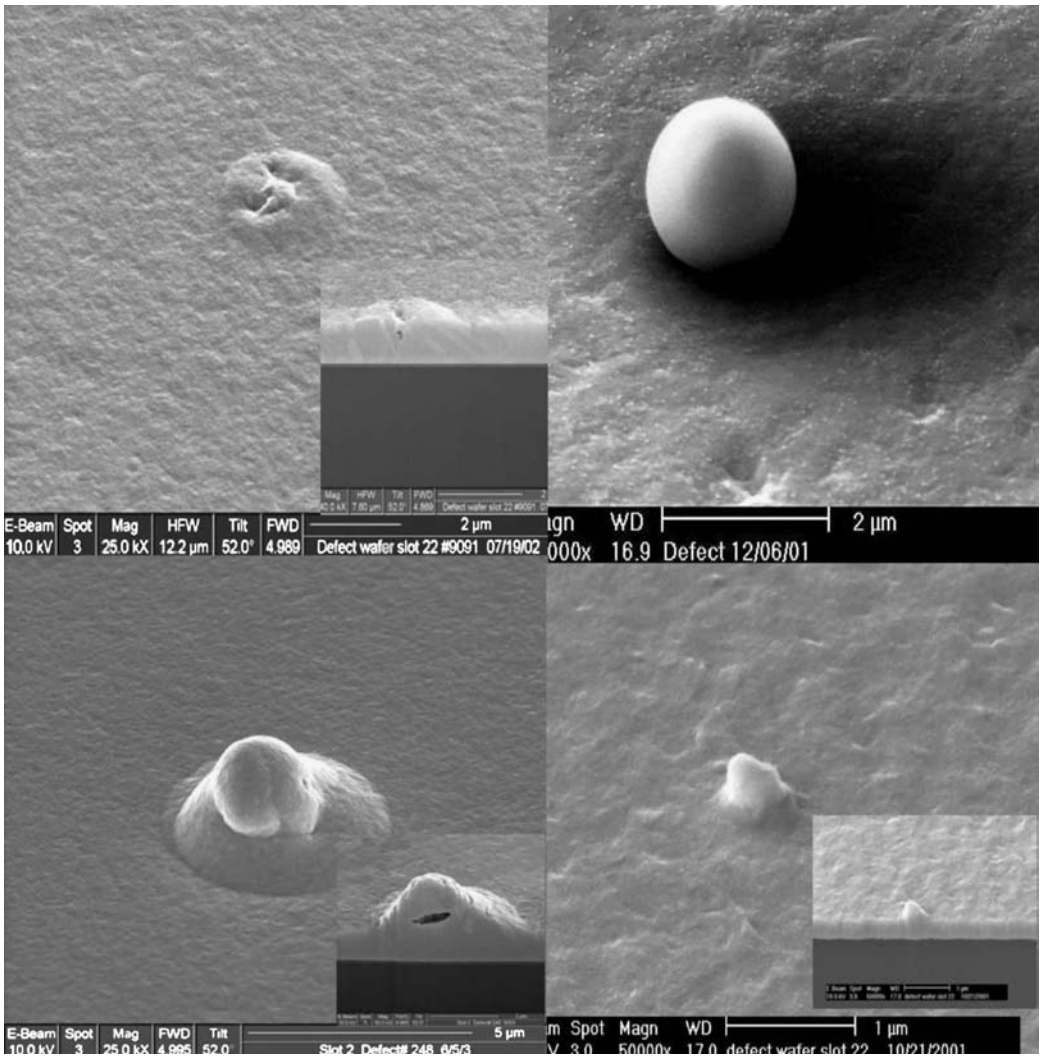


FIGURE 16.30 Abnormal growth defects observed after copper plating.

unfiltered solution from reaching the wafer surface. Examples of irregular copper growth type defects are shown in Figure 16.30.

Particles such as dislodged anode films or extraneous contamination can be incorporated in, or be present on, the plated copper film. To reduce particle defects, the plating solution is normally filtered using 0.05 µm filters immediately before reaching the plating cell and the anode is isolated from the wafer as discussed above. Plating cell material selection and design and adequate filter replacement frequency are also critical for particle free plating performance.

16.5 Modeling Capabilities

The across wafer thickness distribution of copper can be modeled based on plating bath resistivity, plating cell dimensions, seed resistivity, seed type, pattern density, electrical contact pattern, deposition

kinetics, and mass transfer. Several models have been developed which account for some or all of these factors [96–99]. In general, it has been found that seed resistivity, cell dimensions, and bath resistivity are critical in accurately predicting the thickness distribution. Items such as contact position around the perimeter of the wafer and pattern density can be shown to have increasing impact on the thickness distribution as the seed thickness decreases. Mass transfer and deposition kinetics are more difficult to model accurately and are usually less critical in predicting experimental thickness distributions. Mass transfer is relatively complex and usually turbulent due to the combination of wafer rotation and solution flow. This necessitates a full 3D simulation which makes accurate modeling across the full wafer time consuming. Deposition kinetics are influenced by additive adsorption behavior, mass transfer of the additives, and behaviors of the additives on the surface which are not fully understood making modeling based on fundamental principles difficult. Polarization curves at known mass transfer rates can be used to determine an interfacial resistance, which can represent mass transfer and deposition kinetic effects in models.

The fill behavior during damascene plating has been successfully modeled based on surface area loss within features leading to accelerator accumulation [100,101]. In addition, the degree of acceleration capability has been related in models to the degree of hysteresis between anodic and cathodic branches of polarization curves measured for the bath used for filling [101]. Other models have been developed to predict fill based on a lack of suppressor diffusion to the base of high aspect ratio features [102,103]. Although both types of models can predict bottom-up filling, the accelerator accumulation model assumptions are more consistent with the actual chemistries and conditions currently used for damascene filling.

16.6 Process Integration

16.6.1 Feature Profile and Seed Interactions

The performance of the damascene copper plating process is strongly influenced by the characteristics of the incoming PVD seed layers and the etch profile of the underlying dielectric. An ideal seed provides continuous coverage of slightly oxidized copper of uniform thickness within a feature. The physical characteristics of PVD processes, however, tend to result in thicker seed coverage near the top of the feature and thinner coverage near the feature base. This behavior is accentuated when dielectric profiles are re-entrant and result in shadowing of the lower feature sidewalls during seeding. As the seed becomes

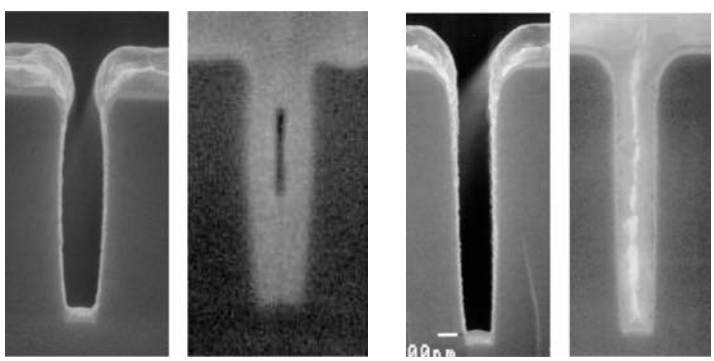


FIGURE 16.31 Profiles of trenches following copper seed layers with excessive overhang and vertical profiles and the resulting fill performance.

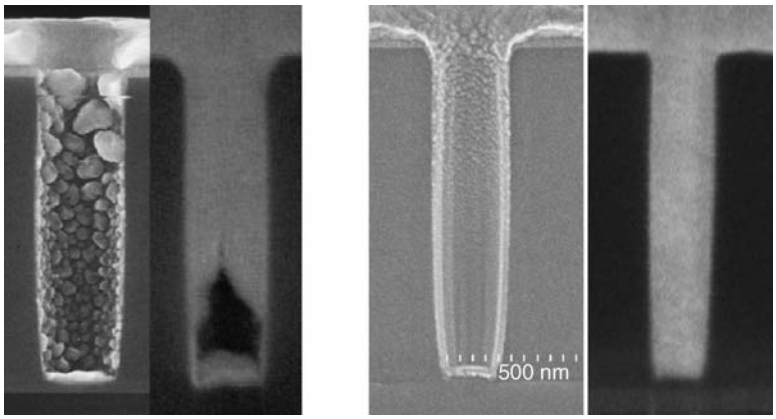


FIGURE 16.32 Topography of copper seed layers on via sidewalls for highly agglomerated and for smooth continuous seeds and the resulting fill performance.

thinner, coverage can become discontinuous or the seed can become fully oxidized, usually near the base of high aspect ratio features [61].

When the seed thickness near the top of the feature becomes too great, the aspect ratio of the remaining open area within the feature increases dramatically and fill without a top center void becomes difficult. Figure 16.31 compares fill in a typical damascene bath for a trench showing significant seed overhang with fill in a higher dielectric aspect ratio trench without overhang. Void free fill is achieved in the feature without seed overhang, while a center void is formed in the trench with excessive seed overhang. Typical pre-plate aspect ratio's which include seed coverage are normally kept under 10 in production to avoid top center void formation.

When seed discontinuity near the feature base occurs due to either agglomeration of the seed copper, because the seed thickness is too low, or due to excessive seed oxidation, bottom voids are observed following plating. Via fill results on an agglomerated seed are shown in Figure 16.32 along with results on a smooth continuous seed layer. Following plating on the agglomerated seed a large bottom void is observed, while deposition on the continuous seed yields void free fill. Seed repair processes utilizing electroless and electrolytic deposition from alkaline solutions were found to be capable of improving fill on discontinuous seed layers [104,105], however, barrier/seed technology improvements have resulted in continuous seed layers with little overhang near the top of the feature. Complete fill of trenches with aspect ratio of up to 20:1 and widths as small as 0.06 μm has been reported in cases where a slight "V" shape exists in the dielectric profile. Improved PVD technology and advancements such as atomic layer deposition result in both an improvement in the seed thickness uniformity within features as well as a reduction in the pre-plate aspect ratio. These developments should allow fill of 32 nm generation or smaller interconnects using existing plating baths, or chemistries modified to more rapidly initiate bottom-up filling.

16.7 Process Control Approaches

In order to maintain consistent process performance, damascene plating equipment normally operates using a large reservoir of plating solution, which is re-circulated through several plating cells. Reservoir volumes of approximately 200 L and plating cell volumes of 5–15 L are typical. The large reservoir volume improves bath stability by acting as a buffer with respect to unexpected dilution, evaporation, or additive consumption variations. During continuous copper deposition, additive chemicals decompose to form by-products, which impact bath performance [106] and result in a finite bath lifetime. To reduce

the impact of additive degradation, plating baths can either be fully replaced prior to a significant change in process performance, or a fraction of the plating solution in the reservoir can regularly be exchanged for fresh electrolyte [107] allowing process operation in a near steady-state condition. An alternative approach, in which the plating solution reservoir is eliminated and small volumes of plating solution are circulated through the plating cells and replaced after a relatively short period of use, has also been implemented [5]. To help maintain process stability, most commercial plating systems use automated dosing of additive components to replace additives consumed during plating.

In all systems, monitoring the concentrations of the plating solution components is required for process verification and correction of concentration deviations. Most damascene plating equipment includes on-line concentration monitoring, and in some cases feedback for dosing of bath components to correct for concentration deviations [108]. Concentrations of inorganic bath components can be monitored using titrations, electrochemical methods [109–111], UV spectroscopy, and other methods. Typical control ranges are less than 2% relative standard deviation. Monitoring of additive components is generally carried out using electrochemical methods such as cyclic voltammetric stripping (CVS) or real time analysis (RTA) [112]. These methods yield analysis accuracy on the order of $\pm 5\%$ relative standard deviation.

Using CVS methods, the voltage is cycled to deposit and strip copper on an electrode in solutions containing bath samples and various standards. The amount of copper stripped during the voltage cycles is measured and related to the concentration of a given additive component. A unique set of bath sample solutions and voltage cycles is normally required for analysis of each additive component, and samples must be discarded after analysis.

Using RTA methods, complex waveforms are applied to an electrode immersed in the plating bath. The current response is measured and related directly to the additive concentration [112]. The plating solution is not consumed during the analysis.

Alternative methods for additive analysis include high performance ion or liquid chromatography [113,114], and mass spectroscopy. These methods provide more direct chemical information such as molecular structure, molecular weight distribution, and inert and active breakdown product concentrations, but are not widely implemented at this time.

Acknowledgments

The data examples presented in various sections of this chapter were collected by numerous Novellus System Inc. employees including Colin Gack, John Sukamto, Seyang Park, and Robert Contolini in support of Novellus Systems Inc. copper electroplating process development and characterization activity.

References

1. Edelstein, D. C., Y. J. Mii, and G. A. Sai-Halasz. "VLSI On- Interconnection Performance Simulations and Measurements." *IBM J. Res. Dev.* 39 (1995): 383.
2. Andricacos, P. C., et al. "Damascene Copper Electroplating for Chip Interconnections." *IBM J. Res. Dev.* 42 (1998): 567.
3. Edelstein, D. C., et al. "Full Copper Wiring in a Sub 0.25 μm CMOS ULSI Technology." Technical Digest, IEE International Electron Devices Meeting. 773. 1997.
4. Andricacos, P. C. "Copper On-Chip Interconnections: A Breakthrough in Electrodeposition to Make Better Chips." *Interface* 8 (1999): 32.
5. Beaudry, C. L., and J. O. Dukovic. "Faraday in the Fab: A Look at Copper Plating for On-Chip Wiring." *Interface* (2004): 40.
6. Misawa, N. et al. "High Performance Planarized CVD–Cu Multi-Interconnection." *VMIC Conf. Proc.* (1993): 353.

7. Nguyen, T., et al. "Electrical Characteristics of CVD Copper Interconnects and Vias." *Electrochem. Soc. Proc.* 97 (1997): 120.
8. Shacham-Diamand, Y., and V. M. Dubin. "Copper Electroless Deposition for ULSI Metallization." *Microelect. Eng.* 33 (1997): 47.
9. Shacham-Diamand, Y., and S. Lopatin. "High Aspect Ratio Quarter-Micron Electroless Copper Integrated Technology." *Microelect. Eng.* 37 (1997): 77.
10. Nguyen, T., L. J. Charneski, and D. R. Evans. "Temperature Dependence of the Morphology of Copper Sputtered Deposited on TiN Coated Substrates." *J. Electrochem. Soc.* 144 (1997): 3634.
11. Shibuki, S., S. Kanao, and T. Akahori. "Copper Film Formation Using Electron Cyclotron Resonance Plasma and Reflow Method." *J. Vac. Sci. Tech. B.* 15 (1997): 60.
12. Schlessinger, M., and M. Panouvic., *Modern Electroplating*. 4th ed, New York: Wiley, 1980.
13. Vereecken, P. M., et al. "The Chemistry of Additives in Damascene Copper Plating." *IBM J. Res. Dev.* 49 (2005): 3.
14. Takahashi, K. M. "Electroplating Copper onto Resistive Barrier Films." *J. Electrochem. Soc.* 147 (2000): 1414.
15. Dukovic, J. O. "Current Distribution and Shape Change in Electrodeposition of Thin Films for Microelectronic Fabrication." *Adv. Echem. Sci. Eng.* 3 (1994): 117.
16. Alkire, R. "Transient Behavior During Electrodeposition onto a Metal Strip of High Ohmic Resistance." *J. Electrochem. Soc.* 118 (1935): 1971.
17. Landau, U., J. J. D'Urso, and D. B. Rear. Electrodeposition chemistry. U.S. Patent 6,113,771, Issued Sept. 5, 2000.
18. Reid, J. D. "Copper Electrodeposition: Principles and Recent Progress." *Jpn. J. Appl. Phys.* 40 (2001): 2650.
19. Reid, J. D., and S. Mayer. "Factors Influencing Fill of IC Features Using Electroplated Copper." In *Materials Research Society Journal 1999 Advanced Metallization Conference Proceedings Red Book Series*, edited by M. Gross, 53–62. Warrenton PA: MRS Press, 2000.
20. Moffat, T. P., D. Wheeler, and D. Josell. "Superfilling and the Curvature Enhances Accelerator Coverage Mechanism." *Interface* (2004): 46.
21. Osterwald, J. "Leveling and Roughening by Inhibitors and Catalysts." *Oberflache-Surface* 17, no. 5 (1976): 89.
22. Bard, A. J., and L. R. Faulkner., *Electrochemical Methods Fundamentals and Applications*. New York: Wiley, 1980.
23. Reid, J. D., and A. P. David. "Impedance Behavior of a Sulfuric Acid—Copper Sulfate/Copper Cathode Interface." *J. Electrochem. Soc.* 134 (1987): 1389.
24. Alkire, R., and D. B. Reiser. "Electrode Shape Change during Deposition onto an Array of Parallel Strips." *Electrochimica. Acta* 28 (1983): 1309.
25. Kessler, T., and R. Alkire. "A Model for Copper Electroplating Multilayer Boards." *J. Electrochem. Soc.* 123 (1976): 990.
26. Dukovic, J. O. "Feature-Scale Simulation of Resist Patterned Electrodeposition." *IBM J. Res. Dev* 37 (1993): 125.
27. Gauvin, W. H., and C. A. Winkler. "The Effect of Chloride Ions on Copper Electrodeposition." *J. Electrochem. Soc.* 99 (1952): 71.
28. Jovic, V. D., and B. M. Jovic. "Copper Electrodeposition from a Copper Acid Bath in the Presence of PEG and NaCl." *J. Serbian. Chem. Soc.* 66 (2001): 935.
29. Healy, J. P., D. Pletcher, and M. Godelough. "The Chemistry of the Additives in an Acid Copper Electroplating Bath, Part 1. Polyethylene Glycol and Chloride Ion." *J. Electroanalyt. Chem.* 338 (1992): 155.
30. Goldbach, S., et al. "Coupled Effects of Chloride Ions and Branched Polypropylene Ether LP-1 on the Electrochemical Deposition of Copper from Sulfate Solution." *Electrochim. Acta* 44 (1998): 323.
31. Yokoi, M., S. Konishi, and T. Hayashi. "Mechanism of the Electrodeposition and Dissolution of Copper in an Acid Copper Sulfate Bath IV: Acceleration Mechanism in Presence of Cl– Ions." *Denki Kagaku* 51 (1983): 460.

32. Tan, M., and J. N. Harb. "Additive Behavior during Copper Electrodeposition in Solutions Containing Cl-, PEG and SPS." *J. Electrochem. Soc.* 150 (2003): 420.
33. Sarma, R. L., and S. Nageswar. "Electrodeposition of Copper in the Presence of 2-Mercaptoethanol." *Surf. Technol.* 12 (1981): 377.
34. Farndon, E. E., F. C. Walsh, and S. A. Campbell. "Effect of Thiourea, Benzotriazole, 4,5-Dithiooctane-1,8-Disulfonic Acid on the Kinetics of Copper Deposition from Dilute Sulphate Solutions." *J. Appl. Electrochem.* 25 (1995): 574.
35. Zhukauskaite, N., and A. Malinauskas. "Mechanism of the Brightening Effect of the Disulfide of Dipropanedisulfonic Acid in Acid Copper Plating Electrolytes." *Protect. Met.* 25 (1989): 132.
36. Moffat, T. B., et al. "Accelerator Aging Effects during Copper Electrodeposition." *Electrochem. Solid State Lett.* 6 (2003): C59.
37. Lichusina, S. B., et al. "Electrodeposition of Cu in Acidic CuSO4 Solutions Containing Accelerating Additive-Sulfite. 4 GALVANOSTATIC Transients for Cu Cathode." *Chemija* 4 (1996): 29.
38. Simkunaite, D., and A. Steponavicius. "Chronopotentiometric Investigation of Acidic CuSO4 Solution Containing Accelerating Additive—DDDS." *Chemija* 3 (1996): 42.
39. Reid, J. D., C. Gack, and S. H. Hearne. "Cathodic Depolarization Effect during Electroplating on Patterned Wafers." *Electrochem. Solid State Lett.* 6 (2003): C26.
40. Frank, A., and A. J. Bard. "The Decomposition of the Sulfonate Additive Sulfopropyl Sulfonate in Acid Copper Electroplating Chemistries." *J. Electrochem. Soc.* 150 (2003): C244.
41. Reid, J. D. "An HPLC Study of Acid Copper Brightener Properties." *Printed Circuit Fabrication* 65 (1987): 65–71.
42. Reid, J. D., and A. P. David. "Effects of Polyethylene Glycol on the Electrochemical Characteristics of Copper Cathodes in an Acid Copper Medium." *Plat. Surf. Finish.* 74 (1987): 66–70.
43. Kelly, J., and A. West. "Copper Deposition in the Presence of Polyethylene Glycol II. Electrochemical Impedance Spectroscopy." *J. Electrochem. Soc.* 45 (2000): 3477.
44. Kelly, J., and A. West. "Copper Deposition in the Presence of Polyethylene Glycol I. Quartz Microbalance Study." *J. Electrochem. Soc.* 145 (1998): 3472.
45. Hebert, K. R., S. Adhikari, and E. Houser. "Chemical Mechanism of Suppression of Electrodeposition by Poly(Ethylene Glycol)." *J. Electrochem. Soc.* 152 (2005): C324.
46. Hope, G., and G. Brown. "A Study of the Adsorption of Polymeric Additives at a Copper Electrode and the Incorporation into Copper Deposits by Electrodeposition." *Electrochem. Soc. Proc.* 96 (1996): 215.
47. Kelly, J., C. Tian, and A. C. West. "Leveling and Microstructure Effects of Additives for Copper Electrodeposition." *J. Electrochem. Soc.* 146 (1999): 2540.
48. Franklin, T. C. "Some Mechanisms of the Action of Additives in Electrodeposition Processes." *Surf. Coat. Tech.* 30 (1987): 415.
49. Jordan, K. G., and C. W. Tobias. "The Effect of Inhibitor Transport on Leveling in Electrodeposition." *J. Electrochem. Soc.* 138 (1991): 1251.
50. Tindall, G. W., and S. Bruckenstein. "Determination of Heterogeneous Equilibrium Constants by Chemical Stripping at Ring Disk Electrode: $Cu + Cu_2 + -2Cu +$ in Sulfuric Acid." *Analytical. Chem.* (1968): 1402.
51. Cheng, X., and B. Hisky. "Fundamental Studies of Copper Anode Passivation during Electrorefining: Part 2 Surface Morphology." *Metall. Mat. Trans. B* 27B (1996): 610.
52. Demedts, G., and A. P. Van Peteghem. "The Corrosion of Copper a Tool in the Investigation of the Reaction Kinetics." *Corros. Sci.* 18 (1978): 1041.
53. Frankel, G. S., et al. "Behavior of Cu(P) and Oxygen Free High Conductivity Cu Anodes under Electrodeposition Conditions." *J. Electrochem. Soc.* 140 (1993): 959.
54. Rashkov, V. S., G. Raichevski, and L. Vuchkov. "Influence of Anodic Films on the Kinetics and Mechanism of the Dissolution of Copper Anodes in Bright Acid Copper Plating." *Bulg. Acad. Sci., Comm. Dept. Chem.* 11 (1978): 459.
55. Reid, J. D., and A. P. David. "Kinetics of Copper Dissolution at Oxygen Free and Phosphorized Anodes." *AICHE Symp.* 83 (1987): 1 (Series 254).

56. Walker, C. T. "Copper Anode Area for High Speed Plating." *PC Fab.* 8–8 (1985): 30.
57. Mayer, S. T., et al. Copper electroplating apparatus. U.S. Patent 6527920, issued Mar. 4, 2003.
58. Ting, C. H., et al. Copper replenishment technique for precision copper plating system. U.S. patent 5997712, issued Dec. 7, 1999.
59. Zhu, M., et al. "Recent Advances in Gap Filling Cu Electroplating Technology." *Electrochem. Soc. Proc.* 99 (1999): 38.
60. Thies, A., et al. "A Novel Electrolyte for Wafer Plating." Materials Research Society Journal 1999 AMC Proceedings Red Book Series. 15,69. Warrenton PA: MRS Press, 2000.
61. Reid, J., et al. "Factors Influencing Damascene Feature Fill Using Copper PVD and Electroplating." *Solid State Technol.* July (2000): 86–103.
62. Moffat, T. P., et al. "Superconformal Electrodeposition of Copper." *Electrochem. Solid-State Lett.* 4 (2001): C26.
63. Josell, D., et al. "A Simple Equation for Predicting Superconformal Electrodeposition in Sub-Micron Trenches." *J. Electrochem. Soc.* 148 (2001): 767.
64. Moffat, T., et al. "Superconformal Electrodeposition of Copper in 500 to 75 nm Features." *J. Electrochem. Soc.* 147 (2000): 4524.
65. Josell, D., et al. "Superconformal Electrodeposition in Sub-Micron Features." *Phys. Rev. Lett.* 87 (2001): 16102.
66. West, A. C. "Theory of Filling High-Aspect Ration Trenches and Vias in Presence of Additives." *J. Electrochem. Soc.* 147 (2000): 227.
67. Georgiadou, M., et al. "Simulation of Shape Evolution during Electrodeposition of Copper in the Presence of Additive." *J. Electrochem. Soc.* 148 (2001): 54.
68. Gill, W. N., D. J. Duquette, and D. Varadarajan. "Mass Transfer Models for the Electrodeposition of Copper with a Buffering Agent." *J. Electrochem. Soc.* 148 (2001): 289.
69. West, A. C., C. C. Cheng, and B. C. Baker. "Pulse Reverse Copper Electrodeposition in High Aspect Ratio Trenches and Vias." *J. Electrochem. Soc.* 145 (1998): 3070.
70. Hayase, M., et al. "Copper Bottom-Up Deposition by Breakdown of PEG–Cl Inhibition." *Electrochem. Solid-State Lett.* 5 (2002): 99.
71. Lee, C., and D. Duquette. "Pulsed Electrodeposition of Copper from Alkaline and Acid Baths for Metallization of Integrated Circuits." *Electrochem. Soc. Proc.* 99-31 (2000): 111.
72. Bason, B. Plating method and apparatus that creates a differential between additive disposed on a top surface and a cavity surface of a workpiece using an external influence. U.S. Patent 6402923, issued June 11, 2002.
73. Mayer, S. T., et al. Method and apparatus for uniform of thin metal seeded wafers using multiple segmented virtual anode sources. U.S. Patent 6773571, issued August 10, 2004.
74. Mayer, S. T., et al. Method and apparatus for uniform electroplating of integrated circuits using a variable field shaping element. U.S. Patent 6402923, issued June 11, 2002.
75. Reid, J. D., et al. Method of electroplating semiconductor wafer using variable currents and mass transfer to obtain uniform plated layer. U.S. Patent 6074544, issued June 13, 2000.
76. Lingk, C., M. Gross, and W. Brown. "X-Ray Diffraction Pole Figure Evidence for (111) Sidewall Texture of Electroplated Cu in Submicron Damascene Trenches." *Appl. Phys. Lett.* 74 (1999): 682.
77. Oshida, Y., P. C. Chen, and J. D. Reid. "Time-Dependent Ductility of Electrodeposited Copper." *J. Electronic Packaging (ASME)* 114 (1992): 448.
78. Harper, J. M., et al. "Mechanisms for Microstructure Evolution in Electroplated Copper Thin Films Near Room Temperature." *J. Appl. Phys.* 86 (1999): 2516.
79. Malhotra, S. P., et al. "Copper Room Temperature Resistance Transients as a Function of Electroplating Parameters." Materials Research Society Journal 1999 AMC Proceedings Red Book Series. 15. Warrenton PA: MRS Press, 2000.
80. Kozlov, V. M. "Influence of Annealing on the Structure and Macrohardness of Electrolytic Copper." *Phys. Met. Metall.* 45 (1979): 184.
81. Sekiguchi, A., et al. "Void Formation by Thermal Stress Concentration at Twin Interfaces in Cu Thin Films." *Appl. Phys. Lett.* 79 (2001): 1.

82. Hu, C. K., et al. "Electromigration in On-Chip Single/Dual Damascene Cu Interconnections." *J. Electrochem. Soc.* 149 (2002): 408.
83. Mallikarjunan, A., S. Sharma, and S. Murarka. "Resistivity of Copper Films at Thicknesses Near the Mean Free Path of Electrons in Copper." *Electrochem. Solid-State Lett.* 3 (2000): 437.
84. Alers, G. B., et al. "Influence of Copper Purity on Microstructure and Electromigration." *2000 IITC Conf. Digest Tech. Pap.* 45 (2000): 2000.
85. Parikh, S., et al. "Defect and Electromigration Characterization of a Two Level Copper Interconnect." *2001 IITC Conf. Digest Tech. Pap.* 183 (2001): 2001.
86. Padhi, D., et al. "Electrodeposition of Copper-Tin Alloy Thin Films for Microelectronic Applications." *Electrochemica. Acta* 48 (2003): 935.
87. Lee, K. H., C. K. Hu, and K. N. Tu. "Insitu Electron Microscope Comparison Studies on Electromigration of Cu and Cu(Sn) Alloys for Advanced Chip Interconnects." *J. Appl. Phys.* 78 (1995): 4428.
88. Aoyagi, M. "Stress-Induced Migration Model Based on Atomic Migration." *J. Mater. Res.* 19 (2004): 2349.
89. Ogawa, E. T., et al. "Stress-Induced Voiding under Vias Connected to Wide Cu Metal Lines." *Int. Reliability Phys. Symp. Proc.* (2002): 312.
90. Alers, G., et al. "Stress Migration and Mechanical Properties of Copper." *Int. Reliability Phys. Symp. Proc.* (2005): 36–111.
91. Shih, C. H. "Design of ECP Additive for 65 nm-node Technology BEOL Reliability." *IITC Conf. Digest Tech. Pap.* 102 (2005): 2005.
92. Varadajaran, S., D. Kalakad, and T. Cacouris. "Understanding and Reducing Copper Defects." *Semicond. Int.* 25 (2002): 125–130.
93. Shaw, J. B., et al. "Voids, Pits, and Copper." *Yield Manage. Solutions* 4 (2002): 8.
94. Lu, J., et al. "Understanding and Eliminating Defects in Electroplated Cu Films." *IITC Conf. Digest Tech. Pap.* 280 (2001): 2001.
95. Lee, D. W., et al. "Effect of Acidity on Defectivity and Film Properties of Electroplated Copper." *J. Electrochem. Soc.* 151 (2004): 204.
96. Deligianni, H., et al. "Model of Wafer Thickness Uniformity in an Electroplating Tool." *Proc. Electrochem. Soc.* 99-1 (1999): 83.
97. Kobayashi, K. "Trench and Via Fill Profile Simulations for Copper Electroplating Process." *IITC Conf. Digest Tech. Pap.* 34 (2000): 2000.
98. Broadbent, E. K., et al. "Experimental and Analytical Study of Seed Layer Resistance for Copper Damascene Electroplating." *J. Vac. Sci. Technol. B* 17 (1999): 2584.
99. Gochberg, L. A. "Modeling of Uniformity and Scale-up in a 300 mm Copper Electroplating Tool." Proceedings of the International Symposium on Fundamental Aspects of Electrochemical Deposition and Dissolution. Electrochemical Society Proceedings, edited by. M. Matlosz, et al., Vol. 99-33, 421–31. Pennington, NJ: Electrochemical Society, 2000.
100. West, A. C., J. D. Reid, and S. T. Mayer. "A Superfilling Model That Predicts Bump Formation." *Electrochem. Solid State Lett.* 4 (2001): C50.
101. Josell, D., et al. "Superconformal Electrodeposition in Sub-Micron Features." *Phys. Rev. Lett.* 87 (2001): 16102.
102. Cao, Y., et al. "Three Additive Model of Superfilling of Copper." *J. Electrochem. Soc.* 148 (2001): 466.
103. Soukane, S., S. Sen, and T. Cale. "Feature Superfilling in Copper Electrochemical Deposition." *J. Electrochem. Soc.* 149 (2002): C74.
104. Webb, E., et al. "Integration of Thin Electroless Copper Films in Copper Interconnect Metallization." *J. Appl. Electrochem.* 34 (2004): 291.
105. Sukamto, J. H., et al. "An Evaluation of Electrolytic Repair of Discontinuous PVD Copper Seed Layers in Damascene Vias." *J. Appl. Electrochem.* 34 (2004): 283.
106. Jiang, Q., R. Mikkola, and B. Carpenter. "Critical Influence of Plating Bath Temperature on Cu Damascene Electrodeposits." *J. Vac. Sci. Technol. B* 17 (1999): 2361.
107. Andricacos, P. C., J. O. Dukovic, and L. T. Romankiw. Method for controlling chemical species concentration. U.S. Patent 5352350, issued Oct. 4, 1994.

108. Contolini, R., et al. "Use of On-Line Chemical Analysis for Copper Electrodeposition." In *Materials Research Society Journal 1999 Advanced Metallization Conference Proceedings Red Book Series*, edited by M. Gross, 117. Warrenton PA: MRS Press, 2000.
109. Freitag, W. O., et al. "Determination of the Individual Additive Components in Acid Copper Plating Baths." *Plat. Surf. Finish.* 70 (1983): 55.
110. Haak, R., R. Teench, and C. Ogden. "Cyclic Voltammetric Stripping Analysis of Acid Copper Sulfate Plating Baths Part 2: Sulfoniumalkanesulfonate-Base Additives." *Plat. Surf. Finish.* 69 (1982): 62.
111. Mansfeld, F. "The Copper Plating Bath Monitor." *Plat. Surf. Finish.* 65 (1978): 60.
112. Wikiel, H. and K. Wikiel. "On-line Monitoring of Copper Interconnect Deposition Processes." AESF SUR/FIN 2000 Proceedings. Orlando, FL: AESF1, 2000.
113. Reid, J. D. "An HPLC Study of Acid Copper Brightener Properties." *Printed Circuit Fabrication* 11 (1987): 65.
114. Heberling, S. S., S. Carson, and D. Campbel. "Monitoring Acid Copper Plating Baths." *Printed Circuit Fabrication* 12 (1989): 72.

17

Chemical–Mechanical Polishing

Gregory B. Shinn
Vincent Korthuis
Texas Instruments, Inc.

Gautum Grover
Cabot Corporation

Simon Fang
United Microelectronics Corporation

Duane S. Boning
Massachusetts Institute of Technology

17.1 Introduction **17**-1
Definitions • History of CMP in the SC Industry • The International Technology Roadmap for Semiconductors and Motivation for CMP • Outline of Chapter Contents

17.2 Equipment and Consumables **17**-6
Classification of CMP Equipment • Pads for CMP • CMP Slurries: Introduction

17.3 Mechanisms and Models **17**-25
Polishing Non-Patterned Wafers • Polishing Patterned Wafers • Chemical Effects in Polishing

17.4 Applications and Issues **17**-42
Dielectric CMP Applications • Metal CMP Applications

17.5 Post-CMP Clean **17**-46
Post-CMP Clean for Dielectrics • Post-CMP Clean for Metals

References **17**-48

17.1 Introduction

The world is not flat. This became apparent with the exploits of Columbus and Magellan at the close of the 15th and the beginning of the 16th centuries. Yet man's initial error in believing the world to be flat is natural. Standing on the edge of the sea or viewing a vast prairie or desert, the earth looks "flat" out to the horizon. This could appear to be the result of natural processes that add silt to lakes, erode mountains, and fill-in valleys—"flattening" the earth. We have coined a new word in English and call such processes "planarization." Of course, man has added to the natural processes. We make surfaces flat for convenience. For example, flat surfaces are easier on which to write than rough surfaces. Structures are easier to erect if you first level the ground. Mankind has constructed tools (e.g., grinders, planes, polishers, and sanders) to make things flat, and he has designed tools to measure flatness. Specialized versions of some of these tools have long been used in the semiconductor (SC) industry to make flat silicon substrates on which to build integrated circuits (ICs). More recently, further developments in chemical–mechanical *planarization* or chemical–mechanical *polishing* (CMP) tools and processing are allowing the SC industry to use multilevel-interconnect (MLI) critical dimensions (CD) of sub-one-tenth micron and to use eight or more levels of interconnect.

An excellent primer for both the new and the experienced CMP process engineer, if they are not familiar with it, is the chapter on "Polishing, Lapping, and Diamond Grinding of Optical Glasses" by Izumitani [1]. That chapter discusses the mechanism of glass polishing and serves as a fundamental basis

for understanding CMP. A detailed summary of glass polishing will be presented in Section 17.3, with the application of these concepts to silicon dioxides used as interlevel dielectrics (ILDs) in MLI structures. We will use Izumitani's definitions of the other two processes—grinding and lapping—that use abrasives to shape and planarize glasses, and we will contrast these with polishing.

17.1.1 Definitions

Grinding is a shaping process that is used to remove material as quickly as possible. The grinding wheel for optical glasses is made by placing 150 bronze pellets, 7 mm in diameter by 1.5 mm high. The pellets are imbedded with diamond grits ranging from 20 to 80 μm in size. Thus, the abrasive is harder than the material to be ground. Water is used mainly as a coolant and to suppress dust. Pressures of about 200 gm/cm^2 are applied to the glass blank with disc speeds of 150 rpm. Depending on the type of glass being ground, material is removed by fracturing or scratching or both fracturing and scratching the surface of the glass. The volume removal rate is high, again depending on glass type and grit size, from 4 to 60 mm^3/min. Thus, the mechanism of material removal is wear.

Lapping is a global planarization process. It is done to remove the surface damage created by the grinding process. Again using the optical glass industry as our example, lapping of the glass blank is usually performed on a rotating, hard, optically flat, cast iron table. Abrasion of the glass is done with either silicon carbide or alpha alumina-grit of size 20–34 μm. These particles are suspended in a liquid and applied to the table as slurry. Pressures of about 100 gm/cm^2 are applied to the glass blank, at table speeds of 60 rpm. Again the solvent, either water or oil, is used as a coolant and to wash away the debris. Material is still removed by fracturing the glass surface. However, the process is self-limiting and the lapping volume removal rate depends on the hardness of the glass, ranging from 0.4 to 2.0 mm^3/min/cm^2. The roughness of the lapped surface typically has 10–20 μm steps.

The *platen* is the bearing surface on which the pad, if there is one, is fixed. It must carry the full down force applied to the work piece by the head or carrier. Originally, the platen was a rigid rotating table that was finished to an optical flatness. Thus, the terms *platen* and table are often used interchangeably in the polishing literature.

Polishing is performed with a rotating table similar to that used to lap. Again, the pressures and table speeds range from 100 to 150 gm/cm^2 and 60 to 200 rpm, respectively. The abrasive for optical glass polishing is typically 10% cerium oxide or 10% alpha alumina in water slurry, with particle sizes ranging from 0.2 to 0.5 μm. Furthermore, the slurry is distributed onto a pad. The pad is either an elastic felt or rubber, or a viscoelastic material such as a pitch-impregnated felt, or a closed-cell molded polyurethane or polyethylene film. If the polishing is done with a dry abrasive or is dispersed in a non-aqueous solvent such as oil or kerosene, the polishing rates decrease by a factor of three and four, respectively, for dry and non-aqueous polishing. Surface roughness approaches the tens of nanometers. The polishing process is viewed as being a chemical softening of the glass surface and the mechanical scraping of this softened layer by the abrasive particle which is imbedded in the pad. The mechanical properties of the pad material affect the removal rate of the glass, which ranges from 100 nm/min to 1 μm/min, depending on pad type and hardness of the glass. Izumitani refers to the pad as the "polisher," signifying the importance he attributes to choose the correct pad material for the glass to be polished.

Planarization is the process by which a surface is flattened. A key requirement is an optically flat work surface on which the lapping or polishing occurs. Figure 17.1 shows schematically a measure of planarity, the step height reduction (SHR) [2]:

$$\mathrm{SHR} = 1 - t_{\mathrm{post}}/t_{\mathrm{pre}} \tag{17.1}$$

where t_{post} is the step height after planarization and t_{pre} is the step height before planarization. An ideal or perfect SHR of 1 results when the post-polish step height is zero. If material is removed equally from "low" or down areas and in "high" or raised areas (e.g., as in a purely chemical wet etch), then no SHR, or SHR = 0, is the result.

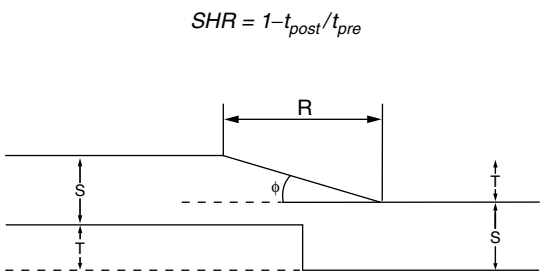


FIGURE 17.1 A measure of planarity, step height reduction. (From Sivaram, S., Monnig, K., Tolles, R., Maury, A., and Leggett, R., *Overview of Planarization by Mechanical Polishing of Interlevel Dielectrics, Symposium on ULSI Science and Technology*, The Electrochemical Society, Pennington, NJ, 1991.)

Global planarization, in SC manufacturing, refers to the flattening of the entire surface of the wafer.

Local planarization refers to the flatness over some reduced area; for SC manufacturing, local planarity usually refers to flatness within a single die or chip.

Damascene processing is a procedure in which the pattern to receive a select metal is first etched in either a host metal or (for interconnect structures) a dielectric. Initially, workers at IBM called this the "recessed metal" (RM) process, as illustrated in Figure 17.2. The term is derived from the jewelry made in old Damascus, Syria, whereby iron or contrasting metal was etched to receive ribbons of precious metals or highlighting softer metals. Usually the inlaid metal was embedded by light hammer taps and the excess

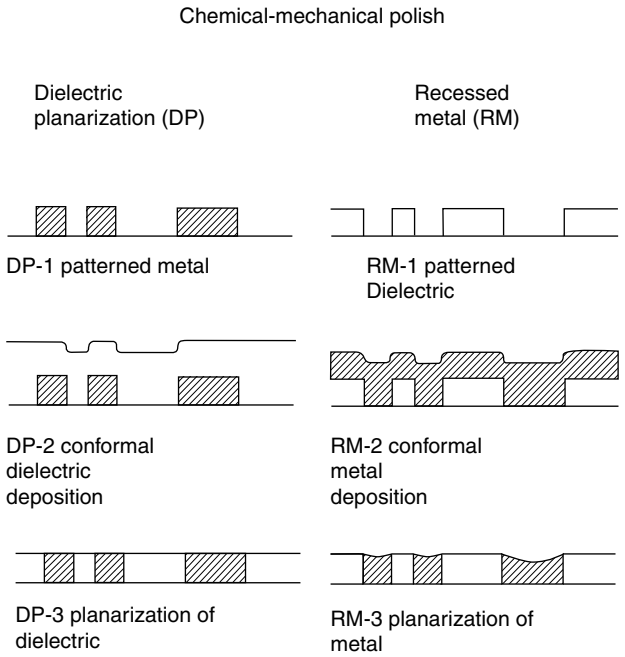


FIGURE 17.2 Comparison of conformal dielectric structure and damascene structure. Processes are dielectric planarization and recessed metal planarization. (From Kaufman, F. B., Thompson, D. B., Broadie, R. E., Jaso, M. A., Guthrie, W. L., Pearson, D. J., and Small, M. B., *J. Electrochem. Soc.*, 138, 3460–5, 1991.)

abraded away to make a striking line of demarcation between the inlaid metal and the host material. In the SC industry, the simultaneous creation of the trench (to hold the metal wire) and the via hole (to hold the metal via which will connect to the underlying wire level), before depositing the conductive layer in the intra metal dielectric (IMD) and the ILD, is called a "*dual damascene process.*"

17.1.2 History of CMP in the SC Industry

Chemical–mechanical planarization or chemical–mechanical polishing as a process for achieving globally planar surfaces originated with the advent of slicing single-crystal wafers from grown single-crystal silicon. The polishing process was done primarily to remove surface damaged layers created by the diamond sawing process, and to achieve a specified wafer thickness and front/back surface coplanarity. The sawing damage extended for several microns into the silicon wafer, and thus the polishing process had to be rapid, yet maintain a planar surface. In the trade-off between high throughput and planarity, substantial wafer material was initially lost. The technology improved as the wafer diameter grew from 25 mm in 1960 to 75 mm in the early 1970s.

The next component of the electronic/computer industry to profit from CMP was the high-density memory disc. Again, CMP was first applied to the planarization of a monolithic material, the substrate on which the magnetic material was to be deposited. Finally, as the structure of these "flying discs" became more complex, CMP was used to re-planarize the surface, which had acquired topography due to masking steps, and multiple depositions of magnetic material and passivation layers, or else the discs "flying characteristics" were compromised [3].

The SC industry introduced the large-scale integrated circuit (LSI), 100 or more gates, in the late 1960s. By the early 1970s, one-thousand-gate circuits were appearing and the random access memory chip was developing. As the complexity continued to grow, one level of metal to interconnect very large-scale integrated circuits was no longer sufficient. Phosphoric-acid-etched aluminum wires doped with copper had become the established and reliable interconnect material technology. Because of its use as a final passivation layer, the industry favored chemical vapor deposited (CVD) silicon oxide from the outset as the ILD for MLI circuitry. However, CVD silicon oxide coatings tended to nucleate preferentially at the top cross-sectional corners of the lower metal layer wires. This would result in a re-entrant fold in the dielectric, and depending on the metal deposition technique, anywhere from a thinned second-level metal to a void would occur in the second-level metal under this oxide outcrop. When the second-level metal was patterned by wet etching, a notch would result in the second-level metal cross-section, which would vary in severity from a shortened mean time to failure to outright open circuits and loss of yield at final test. This double-level metal (DLM) yield problem came to be known as the DLM "edge-effect." At that time, reactive ion etching (RIE) was not available, and sputter deposition of aluminum was still new in manufacturing. Many process modifications were developed to taper the sidewall of the first level metal wire. At the same time, it was realized that a spin-on coating would give a sloped transition from the trench to the top of the first wire. If the coating had high solids content, i.e., low vehicle/solvent content, there would be little shrinkage when the dielectric was cured, and high attenuation of the initial metal step height would be obtained. Spin-on glass and organic resins were the main candidates, with polyimides being the leading resin contender.

Hitachi's R&D laboratories division introduced a polyimide precursor, supplied by Hitachi's Chemical division, known as PIQ. They also developed a wet etch process for vias in the fully cured polyimide, utilizing hydrazine. Due to the health and safety issues associated with hydrazine, this process did not transfer to the U.S. SC industry. Dupont also provided polyimide pre-cursor polyamic acid resins at this time. With the aid of IBM and others in the U.S. SC industry, they developed a large interest in producing this material as a planarizing ILD by the early 1980s, and IBM even introduced a 4K dynamic random access memory (DRAM) product utilizing polyimide as the ILD. However, several reliability issues plagued the polyimides as ILDs. Eventually, RIE became a mainstream manufacturing process and was used to create smaller CD, with controlled slopes, in aluminum wires and the vias of silicon oxide. It was also possible to create a local planarized topography in the ILD, when spin-on glass was applied after

the CVD silicon oxide covered the main part of the wire. Alternatively, sacrificial coating of polymers enabled a non-selective RIE, thereby achieving planarization, by vertically etching back the filled-in surface until all of the coating was gone. Significant improvements in physical vapor deposition permitted good coverage of aluminum in vias with almost vertical walls, $>85°$, $<87°$.

By the late 1980s, sub-one-micron CDs were making aluminum vias unreliable for advanced DRAM products and the industry made CVD tungsten plugs, the standard. By this time, the SC industry roadmap for transistor component and interconnect CDs demanded a new approach to MLI planarity. Workers at IBM had used polyimide as the ILD of the substrate in multi-chip modules. They had globally planarized the polyimide layer by CMP. Also, IBM had a large intellectual property base in high-density memory discs and the use of CMP on these products [3]. The concept of using CMP for multilevel planarization originated at IBM in the early 1980s and had moved to pilot-line production for Logic and DRAM device fabrication by the mid 1980s [4]. Tungsten plugs for vias and contacts were the first use of a damascene process in the industry [5]. Kaufman and coworkers filed a patent for a single damascene process [6]. As the 1990s began, all tungsten plug planarization divided into two technology camps. Those who had excellent RIE etch-back processes in manufacturing at that time tended to remove the excess tungsten metal that way. IBM and companies who had close manufacturing ties to IBM, including Intel and Micron, converted to CMP to remove the excess tungsten. Many SC manufacturers who were building half-micron-node structures converted the ILD planarization from RIE etch-back to CMP, and most found it was required for 0.35 μm technology manufacturing. The companies doing tungsten-plug CMP soon realized significant yield improvements due to the enhanced global and local planarity of the interconnect structure vs. the RIE etch-back technology.

17.1.3 The International Technology Roadmap for Semiconductors and Motivation for CMP

Chemical–mechanical planarization or chemical–mechanical polishing for planarizing shallow trench isolation (STI), post-metal dielectric (PMD), ILD, and damascene structures is an enabling technology for advanced SC technologies. The International Technology Roadmap for Semiconductors highlights the interconnect technology requirements for a number of recent and future device generations [7]. Lithography requirements to pattern the minimum interconnect CD features are one driver forcing minimization of total height variation within the stepper field. A second goal is to minimize total copper line resistance variations due to geometry and scattering effects, which drives the need to have high uniformity, and minimal dishing and erosion of copper features. For example, the goal in the 45 nm technology is to maintain 30% or less metal 1 resistance variability. Copper thickness variation requirements include 8 nm or less thinning (or 10% of the feature height) at minimum pitch due to erosion in 50% pattern density, 500 μm^2 arrays; and 14 nm copper thinning in global wiring due to dishing, in 100 μm wide features. These requirements will continue to force evolution in CMP tool, consumable, and control technologies.

17.1.4 Outline of Chapter Contents

The philosophy of this chapter is to provide a process engineer or process manager with the basic knowledge needed to understand this fast moving technology. Thus, the emphasis will be on core references and articles arising from the height of the development of CMP as a manufacturing technology, with updates where appropriate from the recent literature on new research and development in CMP. Several good books are available to obtain a broad background in CMP. Steigerwald and co-authors cover the topic in their 1997 book; this remains a valuable resource [8]. Other references have also appeared, such as the volume by Li and Miller [9]. A recent collection, edited by Oliver in 2004, provides a valuable update on equipment, pads, slurries, cleaning, and applications [10].

In order to have a physical picture of the CMP process, the equipment and consumables required for the process are presented in Section 17.2. After that, Section 17.3 focuses on mechanisms and modeling,

in which we attempt to provide the process engineer with the basic theory of SC CMP. This is the necessary background to get started. Section 17.4 on applications will allow the process engineer to focus in on a particular process step, such as STI or copper CMP, and understand the core issues and problems. Finally, Section 17.5 deals with the critically important post-CMP cleanup process.

17.2 Equipment and Consumables

Chemical–mechanical planarization or chemical–mechanical polishing equipment has rapidly evolved, from relatively simple tools originally used in virgin silicon wafer polishing, to sophisticated tools adapted for high performance and high throughput polishing, as well as to achieve integration with sensing, control, and post-polish cleaning. While novel polishing tools have been proposed and in some cases brought to market, four polishing types—rotary, linear, orbital, and fixed with orbital carrier—have dominated, as shown in Table 17.1. Additional variants continue to emerge; for example, systems have appeared that are specifically designed for metal polishing which utilize electrochemical reactions to enhance the polish rate and uniformity during the CMP step [11–13].

In the conventional designs, the CMP tool provides a means of applying force to the wafer against the pad, and a means of creating relative motion between the pad and the wafer. The tools or machines that provide the mechanical requirements for polishing are classified as either "lapping" tools or polishing tools. Lapping tools are designed to remove soft material from the work piece surfaces by strictly mechanical abrasion. The lapping tool typically has a hard, inflexible table for a platen, without a pad. It removes material by rapidly impinging a hard abrasive against the work surface. The goal in lapping is to

TABLE 17.1 Classification of Examples Chemical–Mechanical Planarization or Chemical–Mechanical Polishing (CMP) Tools

Vendor and Model	Table Motion-Number Established	Carrier Type-Number Polishers	Built-in Endpoint Methods	Special Features
AMAT Mirra Mesa	Rotary-3	BB–CRRC-4	Eddy Current, Red laser reflectometer	Sequential use of 3 tables. Integrated cleaner
AMAT Reflexion ECMP	Rotary-3	BB–CRRC-4	Eddy Current and laser reflectometry	Platen 1 utilizing electrochemical CMP
AMAT Reflexion Web	Rotary-3	BB–CRRC-4	Eddy current and laser reflectometry	Fixed abrasive technology
Ebara Frex200	Rotary-2	BB–CRRC-2	Visible light	Parallel tool, integrated cleaner
Nikon NPS 2301	Orbital-3	BB–NRRC-3 (Wafer fixed)	Visible light	Down force and slurry through the pad
IPEC 372/472	Rotary-2	FBC/NRRC-1	Absolute temperature	Sequential use of 2 tables
IPEC 676	Orbital-4	BB–CRRC-4	Visible light	Parallel tool
IPEC 776	Orbital-4	BB–CRRC-4	Visible light	Integrated ontrak scrubber
Speedfam V Auriga	Rotary-2	FB–NRRC-5	Visible light reflectometer	
Strasbaugh 6DS-SP	Rotary-1 or 2	FB–NRRC-2		2 carrier size buffing platens, optional 2nd table
Strasbaugh Symphony	Rotary-3	FB–CRRC-4	Visible light reflectometer	
Lam Teres Synergy Integra	Linear-2 integrated cleaner	FB–NRRC-2	Visible light reflectometer	Air bearing platen

achieve a global planarization. In doing so, the surface roughness is high due to the large size of the particles and the hard platen surface. The difference between a lapping tool and a polisher is in the use and the characteristics of the pad on an optically flat platen. Further choices in the consumable materials, i.e., the type, hardness, and average particle size of the abrasive, the solvent and chemical components of the slurry, refine the process for microelectronic CMP. The purpose of this section is to define the elements of a functional design of a polishing tool for SC manufacturing. The section will conclude with a discussion of the key consumables—pads and slurries—used in the process.

17.2.1 Classification of CMP Equipment

The equipment set needed for CMP can be broken down into several subcomponents, with a variety of solutions and designs available for each. In this sub-section, we will first discuss the types of table or platen motions, followed by description of different carrier or head designs. The design of pad conditioners is also important in achieving stable polish performance. Trends in integration of CMP with post-CMP cleaning tools will be summarized. Finally, endpoint methods will be discussed.

17.2.1.1 Table (Platen) Motion

Figure 17.3 is a plan view of a rotary CMP tool, in which the relative motion between the wafer and the pad is created by rotation of both the table and the wafer carrier. The pad is fixed to the optically flat table by a waterproof adhesive. The table and pad combination is sometimes referred to as a "platen," although most authors use table and platen interchangeably. The carrier (also known as the "head" or "quill") is designed to both suspend a wafer for transport purposes, and to apply a uniform down force of the wafer against the pad surface.

The removal rate of surface films, RR, in polishing has been shown to be a function of pressure and speed of the pad surface relative to the wafer surface. In classic early work, Preston claimed that the interaction of these two terms, pressure and relative velocity, are dominant parameters in describing the removal rate dependence on the process [14]. The resulting proportionality constant has been termed the Preston coefficient, K_p, and the simplest model for polishing rate is:

$$\mathrm{RR} = K_p P S_t \tag{17.2}$$

where P is the pressure applied to the wafer and S_t is the table speed at the center of the wafer. The Preston coefficient is an empirical term inside of which there may be complex dependencies on mechanical and chemical effects. It should be noted that there also remains some debate in the literature about the proportionality to pressure and relative speed [15]. In practice, Equation 17.2 as been found effective in approximately describing the removal rate of a homogeneous film on a non-patterned wafer. However,

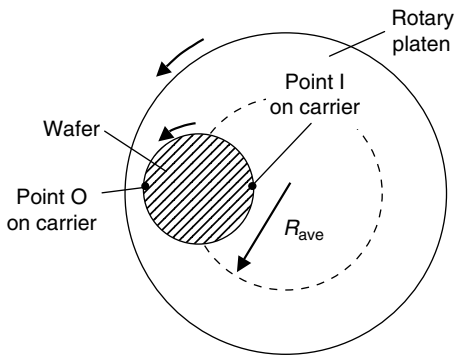


FIGURE 17.3 Schematic plan view of a single carrier-single platen, rotary polisher.

the mechanism by which a wafer with topography is planarized, both globally and locally, requires a more detailed model. That discussion is reserved for Section 17.3. For now, it is sufficient to note that tool pressure and speed are dominant in controlling the removal rate.

With rotary tools, the carrier center of rotation and table center of rotation are offset. When the surface of a wafer is in contact with the platen, if the carrier can rotate freely (as in early glass or virgin wafer polishing machines), the rotation of the table will impart a rotational force to the outside edge, point O, of the carrier greater than to the inside edge, point I, of the carrier. If there is no slippage at the outside edge, the resulting rotational frequency of the carrier will be the same as the table, i.e., they are synchronized [14]. Under synchronous kinematics (including when both the carrier and table are driven at the same speed in the conventional rotary configuration), it can be shown that the instantaneous relative velocity between the pad and the wafer is the same everywhere across the wafer surface. Thus, the relative velocity or speed of travel S_t is most easily calculated at the center point of the carrier (having zero angular velocity), based on the distance from the platen center to wafer center, R_{ave}, and table rotation speed f in rpm:

$$S_t = 2\pi R_{ave} f \tag{17.3}$$

In most CMP tools, the carrier is driven at a specified frequency, in addition to the drive of the table or platen. In these cases, the angular velocities are not necessarily matched. While perfect matching of carrier and platen speeds would appear to be most desirable from a uniformity perspective, in practice, non-uniformity of removal rate is typically observed. One form of non-uniformity is a spoke-like pattern that is due to vibration harmonics. Driving the carrier in rotational synchronization with the table usually causes the worst vibrations. Thus the carrier is usually driven at a constant difference in frequency to the table. If the table was not rotating, but the carrier was rotating, the removal rate would be high on the edge of the wafer and approach zero in the wafer center. Thus, high-speed rotation of the carrier compared to the table rotation will cause "edge fast" non-uniformity to the wafer. In practice, to avoid edge fast polishing (which may also arise due to non-uniformity in slurry flow or pressure application), the carrier is often rotated at somewhat lower frequencies than the table, resulting in a velocity mismatch of a few percent. Since the correct carrier rotational frequency for good circular symmetry on a rotary polisher is not known a priori, taking into account all of these factors, it must be experimentally determined.

As noted earlier, there are at least three other ways to create motion of the pad relative to the wafer surface. The simplest is linear, i.e., the pad is a continuously driven belt against which the wafer is pressed [16]. While uniform relative velocities are in theory achieved with no carrier movement in the linear case, in practice, the carrier is also rotated to average pad/wafer interactions and improve uniformity [16]. Other kinematic arrangements are linear reciprocating, elliptical, or, the special case most used, orbital motions. Reciprocating polishers can be divided into two sub-classes: in most orbital polishing cases, the wafer moves in an orbit about a fixed platen center as shown in Figure 17.4 [17]; in some cases, the platen moves in an orbit around the wafer center [18]. While the kinematics are somewhat different, all are fundamentally similar in achieving a controlled and uniform relative velocity between the wafer and the pad [10]. Most orbital platen polishers have the slurry pumped through holes in a pad that is slightly larger than the wafer. This allows smaller volumes of slurry to be used and the pad area can be smaller, which potentially lowers the overall cost of ownership. Also, many of the orbital systems can generate large relative velocities, which enable these systems to operate at low down force conditions, which is beneficial for the low-k Cu CMP.

Other experimental CMP tool configurations have also been reported [19]. Notable among these are tools in which the pad is smaller than the wafer diameter and is swept across the wafer surface while the wafer is held down by vacuum on an optically flat surface. The velocity of the pad and wafer rotation can be high, allowing low down force polishing conditions. Some potential advantages with this system is the ability to observe clearing non-uniformity and ease of optical endpoint due to the small polishing pad, as well as having the benefits of through pad slurry delivery.

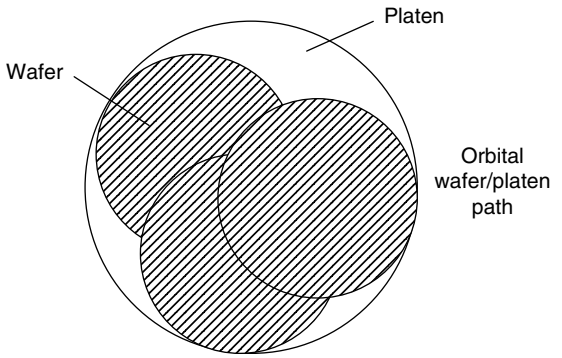


FIGURE 17.4 Schematic of orbital platen motion polisher.

Other new equipment platforms have been introduced for metal polishing which utilize electrochemical techniques [11–13]. With electrochemical polishing, a bias is applied between the surface of the wafer and a cathode to remove conductive material into a surrounding electrolyte. Electrochemical-mechanical polishing occurs when the substrate is still pressed against a pad with relative motion between the wafer and the pad to enhance the removal of material from the surface of the substrate and to achieve planarization. With electrochemical-mechanical polishing, an electrical aspect is introduced to shift the planarization process to the low shear force regime, with removal rate being dominated by the voltage bias instead of down force and table speed. Typically, the electrochemical-mechanical polish is performed to remove the bulk overburden of the copper on the first step, and to minimize topography so that subsequent traditional polishing will have minimal pattern dependent planarization. After the electrochemical-mechanical polish is completed, it is followed by conventional abrasive slurry polishing to clear the overburden, typically <1 K Å, and the barrier films. The pad used for electrochemical-mechanical polishing is unique in that it is designed to have electrodes to set up the bias. With the low down force applied, the pad may have a long life, and an electrolyte is used rather than abrasive-bearing slurries, with potential consumables cost savings.

17.2.1.2 Carrier Design

The design of the wafer carrier is critical to global planarity. Two key design elements are (a) the wafer leveling means and (b) the retaining ring. Classification of carriers depends on which design element is considered primary. The wafer leveling is accomplished by backing the wafer with either a compressible film (a "film-backed carrier" or FBC), a pneumatic bladder (a "bladder-backed carrier" or BBC), or a lip sealed air bearing (an "air-backed carrier" or ABC) [20]. The primary function of the retaining ring is to prevent the wafer from slipping out from under the carrier when high down force and high speeds are used. Wafer retention can be accomplished if the retaining ring extends past the mid-thickness of the wafer. At rest, such a retaining ring might be 100 μm off the pad surface when the wafer is in contact with the pad (referred to here as a "non-contact retaining ring carrier" or NRRC). An alternative retaining ring design causes the retaining ring to contact the pad (referred to here as a "contact retaining ring carrier" or CRRC). In the non-contact NRRC design, the retaining ring is fixed to achieve a preset extension of the wafer surface relative to the ring. The CRRC contacting ring design, on the other hand, incorporates a means of varying the down force on the wide, typically 10–25 mm, bearing surface of the ring. Carriers will be discussed here in historical order.

Figure 17.5 is a cross-section of a typical carrier that uses a perforated pad (also known as a "carrier film") on an optically flat surface plate to carry the wafer [21]. This approach is classified as an FBC–NRRC configuration. The perforations in the pad and flat end plate are used to vacuum attach the wafer during transport to the carrier. The pad is wetted to assure a good vacuum seal. The pad also acts as

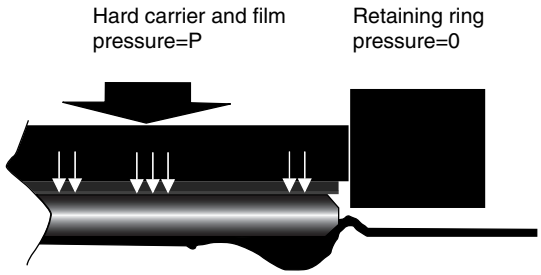


FIGURE 17.5 Schematic of film backed, non-contacting retaining ring carrier (FGB–NRRC). (Drawing courtesy of Applied Materials, Inc.)

a cushion to distribute the down force evenly to the wafer. This is important for local planarization. When the wafer is brought into contact with the table pad surface, the wafer is supported by the platen, and after sufficient down force is applied to the carrier, the vacuum may be released. In order not to slip out from under the carrier, a retaining ring of a hard plastic surrounds the edge of the wafer. The retaining ring is usually made of Delrin™, Teflon™ or related low-wear material. On this type of carrier, the carrier ring protrudes enough to capture the wafer when a down force sufficient to create average pressures of about 2-psi is applied to the carrier. In order not to polish the surface of the retaining ring, it is recessed from the surface of the wafer by about 100–200 μm. Since the wafer is held rigidly to the flat surface of the carrier, there must be a means of making the wafer surface co-planar with the surface of the platen. This is accomplished by a universal joint known as a gimbal.

Figure 17.6 is a cross-section of a "bladder backed carrier" [22]. This carrier substitutes a rubber inner tube style bladder for the soft surface pad of the FBC. A flat, rigid perforated frame is sealed inside the bladder. Thus the chamber created by this bladder-perforated frame can be alternately vacuumed or pressurized. When the wet backside of a wafer is placed against the bladder and the chamber is vacuumed, the bladder conforms to the holes in the perforated frame and creates small suction cups that hold the wafer for transport. In principle, both types of retaining rings are possible with the BBC: fixed or adjustable. Because down force is created by pressurizing the bladder, larger vertical displacement of the wafer is possible. Thus in practice, the contacting retaining ring is also adjustable by a separate pneumatic control. When the carrier brings the wafer into contact with the platen, first the pressure is applied to the retaining ring bladder with a sufficient down force to clamp the wafer within the retaining ring. Next, the

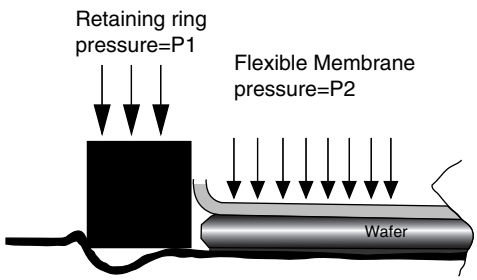


FIGURE 17.6 Schematic of bladder backed, contact retaining ring carrier (BB–CRRC). (Drawing courtesy of Applied Materials, Inc.)

wafer bladder may be vented or pressurized. Usually the platen is stationary when the carrier retaining ring makes contact and the wafer bladder is vented until the platen comes to speed. Then the wafer bladder is pressurized and the carrier brought to speed. Since the wafer "floats" on the bladder bubble, global and local planarity of the wafer to the platen surface is accomplished without a mechanical gimbal.

The ABC [20] functions very much like the bladder backed carrier. Instead of a bladder to contain the air pressure or vacuum behind the wafer, it has a compliant lip seal around the backside edge of the wafer. The lip seal is several millimeters back from the edge of the wafer. In order to self-regulate the backing air pressure, both the lip seal and the contact retaining ring must permit air pressure to leak in a controlled manner. This ABC–CRRC design allows slurry particles to accumulate in the lip seal to retain ring cavity. This causes back of the wafer edge contamination, difficulty in clean up, and potential for lithography yield loss [20].

Both the FBC–NRRC and the BBC–CRRC can be used with any platen style tools. Although uniformity of film removal is dependent on many factors, the dominant factor is the carrier type and the means used to control the shaping of local forces. Minimization of concentric non-uniformity, i.e., center fast vs. edge fast polishing, is the first adjustment to make. Edge fast polishing is a common condition. As previously mentioned, this can be caused by the carrier speed being too fast, poor slurry transport to the center of the wafer, or too high edge pressure on the wafer. Slowing down the carrier speed and increasing the slurry flow will correct the first two causes. In the case of high edge pressure, the center of the wafer or carrier must be shaped to increase the pressure at the center of the wafer. In the case of the FBC, a convex wafer shape is obtained by applying pressure to the back of the wafer. The convex shape of the wafer can be enhanced by restricting pressure ports to the center of the wafer or by blocking ports at the edge of the wafer. The CRRC corrects this problem by applying a higher pressure to the retaining ring. This action causes the high polishing rate to be shifted out to the wide retaining ring, resulting in more uniform edge to center pressure on the wafer; see Section 17.3 for a discussion of another interpretation of the "wafer edge effect." This is a very elementary discussion of the carrier mechanical adjustments that are available to the process engineer to compensate for wafer edge to center non-uniformity. Another typical non-uniformity pattern is edge-slow–center-slow with a fast "donut" region between. Practical adjustments to address this problem include lowering the pressure and raising the platen speed.

Several recent wafer carrier designs have introduced means for controlling and localizing the applied pressure to some number of separate radial "rings" or "zones" on the wafer [23]. The resulting pressure regions on the wafer typically overlap somewhat, but provide a powerful means to tailor and optimize the polishing uniformity across the wafer.

17.2.1.3 Design of Pad Conditioners

An important tool design factor is polish by-product removal. The slurry and pad must promote both film polishing and abrasion product removal. The price of closed cell polyurethane pads is high, \$0.06–\$0.10/cm^2. Since a 200 mm wafer requires a 300 mm diameter pad for an orbital table and a 510–550 mm diameter pad for a conventional rotary table, the initial cost for the pads can be \$42 and \$125 each, respectively. Thus to achieve pad cost of ownership of \$0.05/wafer or less, the smaller pad must polish 800 wafers and the larger pad must polish 2500 wafers before they are discarded. These are challenging numbers to achieve in practice.

The most widely used pads are closed cell polyurethane. As these pads polish successive wafers, there corresponding flattering of the polyurethane asperities on the surface of the pad, as well as a build up of smaller than average abrasive particles and waste particles from the polished film in the closed cell pores. This "loading" of the pad cell pores lowers the polishing rate. Also, large waste particles can become embedded in the pad and cause scratches. For these reasons, periodic or continuous conditioning of the pad surface is required.

The simplest form of conditioning was originally done with a nylon-toothed brush. This is an adequate conditioning for large particle removal or conditioning soft, woven, or high nap pads. For conditioning the harder, molded, closed-cell polyurethane pad materials, e.g., the Rohm and Haas IC-1000 or IC-1400

series, diamond conditioners are required. The most common conditioning tool is a disc with 80–120 mesh size diamonds embedded in a nickel plate. These discs come in an open screen form [24], or a solid disc form with a ridge and valley pattern [25]. Slurry particles can build up in the holes of the screen style discs. If these agglomerated particles dislodge onto the pad, while polishing, wafer scratching will result. Periodic cleaning of the screen disc avoids this problem. Both the discs are of iron core material and are clipped on to magnets in the head of the conditioning arm. The key to the effectiveness of these diamond discs is the adhesion of the industrial diamonds. Diamonds are coated better on the conditioning discs with brazes that wets them than entrapping them with a plated metal [25]. However, with extended use, the diamonds tend to become dislodged. Since 80 grit diamonds are 100–200 μm in diameters; a significant scratch pattern develops, the longer they are retained on the pad. A new disc is expected to last for several pads and 10,000 or more wafers. The useful life of these discs is judged, in practice, by the removal rate of oxide dielectric films, under standard conditions, with a new pad. Visual inspection of the diamond content is also possible. When the diamond density falls to less than a quarter or a third of the original content the discs are generally discarded of refurbished. The retention of diamonds is generally good in neutral and basic slurry conditions. Under acid conditions, particularly at pH below 2, the nickel plate film is etched and the diamond loss rate increases. With acid slurries, ex situ diamond disc conditioning with de-ionized water is preferred. The ex situ conditioning step lowers the throughput of metal polishing with acid slurries.

In order to overcome the disadvantages of the diamond disc conditioner, a single diamond point conditioner has been proposed [26]. Since a single large diamond point is more firmly fixed to the head of a conditioning arm, and since its loss is more easily discerned, problems of scratching are eliminated and maintenance is minimized. The single point is a diamond turning tool. Thus it will cut a groove in the pad surface and for this reason the tool design limits the tip penetration.

For manufacturing operations that do not have access to this proprietary conditioning tool, the preferred surface for a Rohm and Haas IC-1000 pad is a K-groove. These grooves have been diamond point turned into the flat molded surface of the IC-1000 material during pad manufacture. The grooves give the pad more slurry carrying capacity. However, the polishing work is done on the ridges, between the grooves. It is the cell pores of the ridges that need to be cleaned and the surface of the ridge that needs to be rejuvenated. Many 80 grit diamonds on a solid or screen disc are more effective in conditioning this surface contour than a single point diamond and is the preferred conditioning tool for the grooved or perforated pads. Dislodged 80 grit diamonds tend to be collected in the deep K-groove, rather than the ridge top. This collection of the diamonds in the grooves minimizes scratches on the wafer surface. However, the pad should be cleaned between wafers with a high-pressure spray that is designed to free the pad of these dislodged diamonds. This high-pressure spray is more critical at end of the pad life, when the groove has become shallow, and the freed diamonds can fill the groove and scratch the wafer.

17.2.1.4 Post CMP Cleaning—Polish Tool Integration Issues

Post CMP cleaning is a critical step. It must satisfy several integration requirements. Abrasive particles, pad, and wafer film debris must be removed. The surface must be dried and, in the case of inlaid metal structures, corrosion must be controlled or avoided. These defect types can be seen in Figure 17.7. In one example of tool configuration, rotary action in both the horizontal and vertical axes allows the arm of the 472 tool to pick up wafers from a dry load station [27]. The wafer has been delivered from the input cassette to the load station by a horizontal transfer mechanism. The arm then positions the carrier over the primary polish table, brings the wafer into pad contact and applies down force and rotates the carrier. At the conclusion of the polish step, the arm rotates up and over to the secondary table, rotates down and again applies down force to buff the wafer. At the conclusion of the buff step, the arm picks up the carrier and rotates back to the unload station where the wafer is released to the water track which transports the wafer to the unload cassette. This is a very compact tool. However, with polish times of 2 min per wafer and a 30-s buffer step, throughput can never surpass 20 wafers/h. In order to get a throughput of 40 wafers/h or more, two or more wafers must be polished and cleaned simultaneously. This requires the automation of the CMP tool to have multiple robots; at least one for wafer transfer from and to the cassettes and one to apply

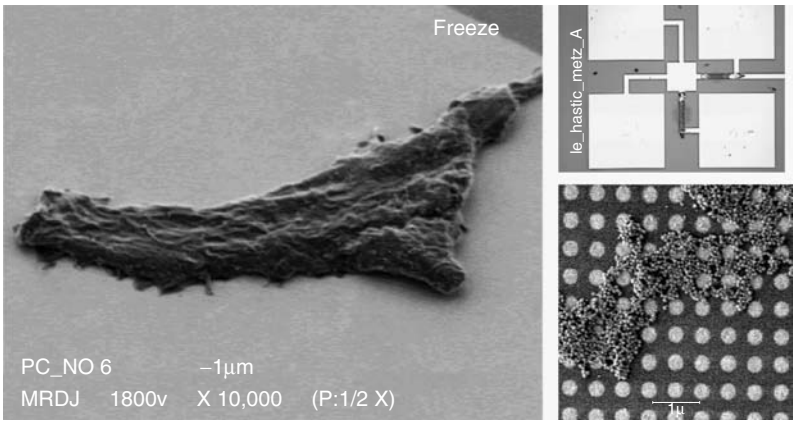


FIGURE 17.7 Typical defects in chemical-mechanical polishing (CMP).

down force and otherwise control the carrier operations. The latter robot is usually controlling two or more carriers and it is given a unique name like "head," "bridge," "cross" etc. In Table 17.1, the complexity of the handling system is implicit with the increasing number of carriers and tables.

Two equipment approaches have been implemented for the clean-up step. The dominant machine is a cassette-to-cassette track with stations for double-sided scrubbing, rinsing, and spin-drying. Since the CMP process and this cleaning process are single wafer-at-a-time processes, there is an advantage to combine the two processes into a single integrated machine. By joining "off-the-shelf" double-sided scrubbers, through use of a wafer transfer robot, CMP equipment manufacturers could achieve this industry equipment standard. However, cleaning processes are still evolving. What appears to provide adequate cleaning of the traditional oxide dielectric or present tungsten plug CMP process may not be suitable for Cu–low-k applications of the future. The chemical hood approach provides a wider range of chemistries and better temperature control than the double-sided scrubber does. Megasonic agitation is substituted for mechanical brushing to remove particles by this strictly chemical means [28]. The traditional wet chemical hood processes cassettes of wafers. Because abrasive particles are harder to remove once they have dried on a wafer surface, wafers must be collected and maintained wet, before they are transferred to a chemical cleaning hood. Other than the large floor space and chemical support required, with 200 mm-equipment, this is not a disadvantage to this clean-up process. However, the 300 mm-equipment standard poses some problems for the integration of a megasonic tank style cleaning system. With megasonic energy, sophisticated chemistry, and accurate temperature control all required, 300 mm CMP equipment is tending to integrate these functions into the tool.

17.2.1.5 Endpoint Methods

In order to achieve total factory automation, in situ control of the process is required. Thus, a final critical equipment component for the CMP process is a method of detecting the so-called "endpoint" of the process. Two types of processes need to be controlled: (a) polishing a dielectric to a known final film thickness and (b) polishing an inlaid, usually multi-layered metal film and stopping on the surrounding dielectric, after all the intra-inlaid metal is removed. Several in situ endpoint methods exist for polishing inlaid structures [29–32]. No satisfactory in situ endpoint method exists for polishing homogeneous dielectric films to a known final thickness. It is instructive to consider why this is so difficult as an instrumentation goal.

Figure 17.8 shows the typical cross-section of a conformal dielectric coated on an etch-patterned aluminum wire, before polishing. A known thickness of the dielectric, including an excess consistent with the planarization ratio of the CMP process, has been deposited, Z_0. For device performance reasons

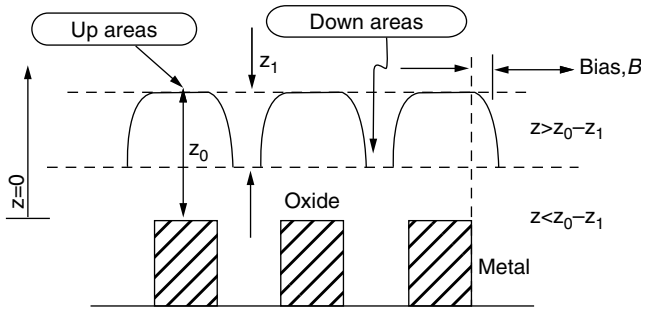


FIGURE 17.8 Cross section of conformal dielectric on pattern metal wire pattern. (From Ouma, D., Modeling of Chemical Mechanical Polishing for Dielectric Planarization, Ph.D. thesis, MIT, Cambridge, MA, 1998.)

it s important to remove just this excess and stop at the desired dielectric thickness, Z, above the top of the metal wire, such that, $Z > Z_0 - Z_1$. The most direct way to measure Z_0 and Z is by some optical methods, e.g., ellipsometry, laser reflectometry, visible wavelength scanning spectrometry, etc. Spot sizes on all of these thin film dielectric-measuring tools are small enough to use a 100 μm^2 pad as the reference metal surface. However, none of them are capable of reliably focusing on the exact same pad if the wafer is revolving on a carrier. The mechanical tolerances of the tool will not allow the light beam to shine exactly on the same pad at each sampling opportunity. Once the MLI becomes two or more levels high, scattering from lower layers confounds optical responses. Thus, the best endpoint method for dielectric thickness control is an inline control method such as that described by Nova Instruments, in which a trial wafer is polished for 90% of the standard time, measured, and a corrected time is applied to the third wafer in lot [33]. Data from the second wafer, polished as the first is being measured, either confirms the result of the first wafer or is used to correct the time for the fourth wafer. If the process is stable, only the first two wafers will require touch-up polishing. If the removal rate is increasing or decreasing in a constant manner as the lot is polished, the feed forward nature of the control system can compensate, if the first two wafers correctly predict this trend. Using an automated inline control method is faster and more accurate than the present manual off-line methodology used by the industry.

Many more endpoint methods are available for inlaid structures. One can visually detect, with the aid of a 100–500× microscope, the appearance of the clearing of the metal film. Thus a simple laser reflectometry tool has become a useful endpoint sensor for inlaid metal polishing (Figure 17.9) [34]. A red laser light source and detector are placed in the table and the beam is projected, through a window in the pad, at 45 degrees to the wafer surface and collected at 45 degrees from the surface. A typical reflectance vs. polish time curve is shown in Figure 17.10 for polishing a copper film with a tantalum barrier layer inlaid on oxide dielectric [34]. After the maximum reflectance is reached, the process starts to break through the copper layer to the semitransparent tantalum layer, and a sharp drop in reflectance results. The sharpness of the transition of the "S" shaped curve depends on how uniform the clearing process is. A small step is seen as the weakly reflecting barrier layer is cleared and the light is scattered by the break through to the dielectric surface. The copper film, filling the trenches, continues to reflect a small constant amount of light. A patent has been issued on this method [30]. Two additional properties easily detected as a function of work done are heat flux or power required. As the polish process transitions from a homogeneous layer of one hardness to a heterogeneous layer, where most of the new layer is either harder or softer than the homogeneous layer, a change in work to be done by the polishing tool occurs. In another approach, two optical thermo-sensors are used, one sensing the pad before, and the other after, the carrier to provide a differential temperature curve [29]. A patent and a practical instrument have resulted from sensing the drive current of either the carrier motor or the table motor [31]. The difference in sound transmission of the films that make up composite dielectrics and metal vs. dielectric has led to a patent on sensing acoustical waves during polishing [32].

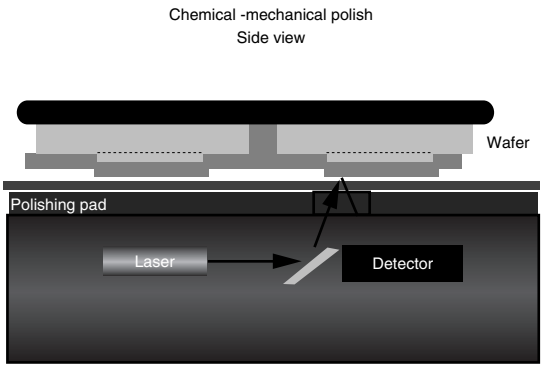


FIGURE 17.9 Schematic of AMAT's *Mirra* laser reflectometer light path. (Figure courtesy of Applied Materials, Inc.)

17.2.2 Pads for CMP

The polishing pad is known to be one of the most critical elements in a CMP system, with substantial impact on polishing rate, planarization, uniformity, defectivity, and other process results. Steigerwald [8] discusses pad mechanical properties and surface roughness, as related to removal rate and planarization. A number of models have been developed to relate the viscoelastic properties of soft conformal pads and stiffer but flexible compressible pads, through a mechanistic relationship of polishing pressure distribution, to the resulting planarization. However, a great deal of the knowledge related to pads and pad use remains empirical; in this section we focus on such empirical observations about pads and their effects during polishing.

In principle any hydrophilic, woven, non-woven or semi-porous fabric or viscoelastic material could be used as a polish pad. In practice, closed cell, cast polyurethane with a filler material to control hardness or polyurethane impregnated polyethylene felts are the materials of choice [35]. Table 17.2 lists the important properties of five polishing pads (manufactured by Rohm and Haas) [36].

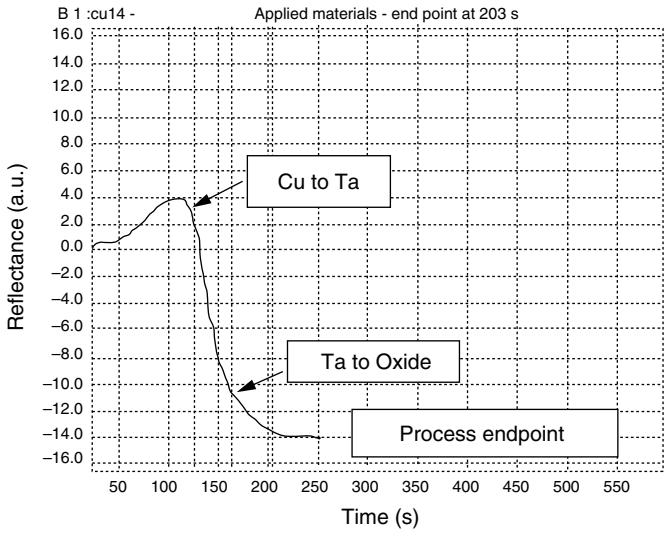


FIGURE 17.10 Typical reflectance curve for polishing Cu/Ta/oxide film. (Drawing courtesy of Applied Materials, Inc.)

TABLE 17.2 Rohm & Haas Polyurethane Pad Properties

Pad	Specific Gravity	Compressibility (%)	Hardness	Type
Politex Supreme	Not Available	High	Soft	Micro Porous Felt
Suba IV	0.3	15	61 (Shore A)	Urethane impregnated polyester felt
Suba 500	0.34	7	72 (Shore A)	Urethane impregnated polyester felt
IC-1000	0.63–0.85	2.25	57 (Shore D)	Rigid micro porous cast polyurethane
IC-1400	Top layer IC-1000, bottom layer foam	0.7–7	Flexible	Equal to stacked IC-1000/Suba IV

Source: From Rohm & Haas Corporation, 451 Bellevue Road, Newark, DE 19713.

The pads designated "Suba" are polyurethane impregnated polyethylene felts. The "IC" designation is reserved for micro-porous, filled, and cast polyurethane. Due to the basic fiber structure of felts, the pad made from this material is more porous and can carry a larger volume of water or slurry. This property is reflected in the low specific gravity of these materials, <0.4 g/cc, and the greater compressibility, $>7\%$. The Suba pads are classified as soft polish pads. The IC-1000 is a filled, cast pad and its low compressibility, $<5\%$, and higher specific gravity, >0.6 g/cc, classifies it as a hard polish pad. IC-1400 is a hybrid. Felt pads have been observed to provide faster removal rates. This may be due to the more open structure of felts, providing higher slurry carrying capacity and better transport of slurry to the wafer surface and transport away from the wafer surface of polish debris. However, the IC-1000, either perforated or grooved, provides reasonable slurry transport; but mainly it provides lower within-die non-uniformity and hence better planarization. Process engineers like the better global planarization of the IC-1000. However, in some applications it does not conform to the wafer surface as well. Thus engineers in the field began stacking the IC-1000 on a Suba IV pad. The IC-1400 is a manufactured solution to these needs: it is a single molded pad, consisting of an IC-1000 pad molded to a more porous backing pad that simulates the compressibility of the Suba IV pad. Finally, the Politex Supreme pad is classified as a finishing pad. It is used to buff the wafers with water and other chemicals to remove the particles and reduce scratch defects.

There was a time when Suba 500 was preferred for the tungsten plug polishing process. The felt pad provided the higher removal rate necessary for productivity with the earlier slurries. With 0.5 μm and larger CDs, the planarity was not critical. As the industry has pushed below 0.5 μm CDs, planarity is an important issue. Fortunately, slurry vendors have responded with faster polish rates and the IC-1000 type pad surface can be used both for dielectrics and tungsten. Rapid copper damascene polishing also has been demonstrated on the IC-1000 pad surface.

At the present time, Rohm and Haas is the dominant supplier of CMP pads. Unfortunately, there is no simple physical property of a polish pad that is known to lead to good quality for all CMP tools and consumable sets. Typically, IC manufacturers will specify an allowable range for specific gravity for the rigid micro-porous cast pads (i.e., IC-1000), and other properties.

17.2.3 CMP Slurries: Introduction

Slurries are the third critical component of the CMP process. The choice of slurry can have a significant effect on integration, manufacturing, logistics, reliability, and yield. CMP users have seen, for example, that the choice of W CMP slurry can have a significant effect on device yield [37,38].

As CMP technology has evolved, slurry technology had to evolve to keep pace with the demanding requirements of the industry. A significant amount of intellectual property has been invested in slurry

formulation, both by the supplier base and the IC manufacturers themselves [39]. Today's CMP slurries are increasingly tailored for specific process applications (like STI or W local interconnects) as well as specific polishing tools and other consumables like pads. Chemical–mechanical planarization or chemical–mechanical polishing slurry manufacturing requires expertise in the areas of particle synthesis; dispersion, mixing, and filtration; electrochemistry; colloid science and surface chemistry; fluid dynamics; and numerical analysis [40]. Operations expertise is crucial for a supplier to flawlessly manufacture the slurry day in, day out, and deliver it to a factory for high volume manufacturing.

17.2.3.1 Slurry Compositions

Most CMP slurries consist of an abrasive dispersed in water and mixed with reagent chemistry: the latter may include oxidizers, surfactants, complexing agents, buffers, and/or other additives used to impart specific selectivity to films (for example, a W: Ti or W: TEOS selectivity). For metal CMP slurries, the oxidizer is often packaged as a separate component. The dispersed abrasive and the oxidizer are then mixed either at the point of use (POU) or in a mix tank before being delivered to the polisher. The two components are often kept separate for reasons of physical or chemical stability.

17.2.3.1.1 Abrasives

First generation slurries developed for polishing oxide substrates consisted of fumed silica abrasives, while the first generation metal CMP slurries used alumina as the abrasive [41]. Typically slurries can contain about 1%–35% abrasive dispersed in water. Concentrate slurries are popular for oxide CMP since they contribute to lower cost of ownership (CoO). These are shipped at high solids concentrations and then diluted on-site for use. In such cases, however, it is extremely important to exercise care in the dilution process and follow the manufacturer's recommendations.

The main types of silica abrasives used for CMP are fumed and colloidal, with fumed silica slurries occupying a far larger portion of market share. Fumed silica is typically manufactured at a high temperature in a flame by the oxidation of silicon tetrachloride. Advancements in burner technology over the last decade have allowed suppliers to deliver dramatically improved abrasives to the CMP end user. These improvements have delivered lower defectivity and more consistent performance. Abrasives are now tailored for specific CMP applications by changing the physical properties and the surface chemistry of the abrasive. In addition, colloidal silica slurries have also entered the market as an alternative to fumed silica. Colloidal silica is manufactured by hydrolysis of sodium or other silicates. This is followed by an ion exchange to reduce the sodium levels for IC applications. All silica manufactured in this manner is amorphous. Figure 17.11 shows SEMs of typical fumed and colloidal silica particles. Oxide slurries containing fumed silica generally produce higher removal rates for the same level of solids than slurries with colloidal silica.

Alumina has traditionally been used as an abrasive for metal CMP slurries, both because it has resulted in higher rates and higher selectivity to the ILD layer. This alumina may either be manufactured by a fumed process similar to silica or by precipitation from aluminum hydroxide. After manufacture, the alumina undergoes a calcination process that transforms the phase of the abrasive. The alumina may then be a mixture of more than one phase (α, β, γ, or δ), depending on the temperature of calcination. Research has shown that the phase and/or hardness of the alumina may have an effect on the polishing rate as well as on the defectivity of the surface [42,43]. Typically, α alumina particles tend to be larger, harder, and denser, giving both the highest rates as well as the highest defectivity. Figure 17.12 shows typical phase transition temperatures for alumina along with the densities. Defectivity issues with alumina have resulted in a move toward silica for metal CMP, with some slurry achieving both high rates and high selectivity to ILD. Figure 17.13 shows a TEM of alumina particles manufactured by a high temperature process.

In addition to silica and alumina, the past few years have seen the emergence of other abrasives for CMP. Ceria slurries have been formulated for oxide and STI CMP, although ceria cannot be easily manufactured at small particle sizes. A self-stopping process for oxide CMP has been demonstrated with ceria abrasive [44]. In addition, manganese dioxide slurries have been tried for metal CMP, with the

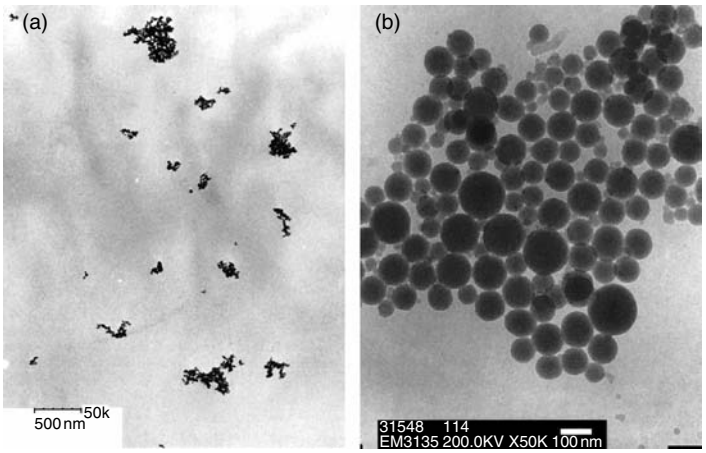


FIGURE 17.11 SEM of typical fumed (left) and colloidal right (silica). Note that scales are different for the two figures.

manganese dioxide playing the role of both abrasive and oxidizer [45,46]. Zirconia slurries have also been developed for CMP, although they are yet to gain a widespread acceptance among end users.

17.2.3.1.2 Chemistries

In addition to the abrasive, slurry chemistry plays a significant role in determining the slurry performance. Additives for slurries may consist of oxidizers, buffers, stabilizers, surfactants, passivating agents, complexing agents, corrosion inhibitors, or other agents for imparting selectivity to various films.

Typically, slurries for oxide CMP consist of few chemical additives, being comprised mainly of an abrasive. In addition to the abrasive, slurries may contain additives for colloidal stability and/or buffers to withstand pH shock. Oxide slurries are also available in concentrate form at 25% or higher solids; they can then be diluted for use. Ceria slurry containing high concentrations of surfactants has also been demonstrated.

Metal CMP slurries often have more constituents than oxide slurries. The oxidizer is often separated from the abrasive, as in the case of ferric nitrate, for reasons of colloidal stability or shelf life. If the oxidizer is packaged with the abrasive (as one component), the colloidal stability and/or chemical stability of the resulting mix could be adversely affected. Common oxidizers include potassium ferricyanide, ferric nitrate, hydrogen peroxide, and potassium iodate [47]. Various papers have compared the performance of slurries containing these oxidizers [1,2,48].

Abrasives may be suspended in solution by either electrostatic (charge) or steric stabilization (Figure 17.14). If electrostatic stabilization is used, the zeta potential (ZP) curve of the abrasive becomes important in determining regions of stability; it is important to have high ZPs at the pH of the slurry. If steric stabilization is used, however, the stabilizing molecule (often a dispersant like a polymer or surfactant) must be chosen so as not to affect any of the other properties of the slurry [49].

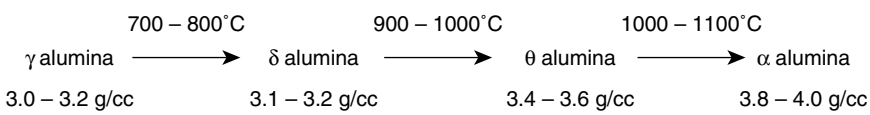


FIGURE 17.12 Phase transition temperatures and densities for different phases of alumina.

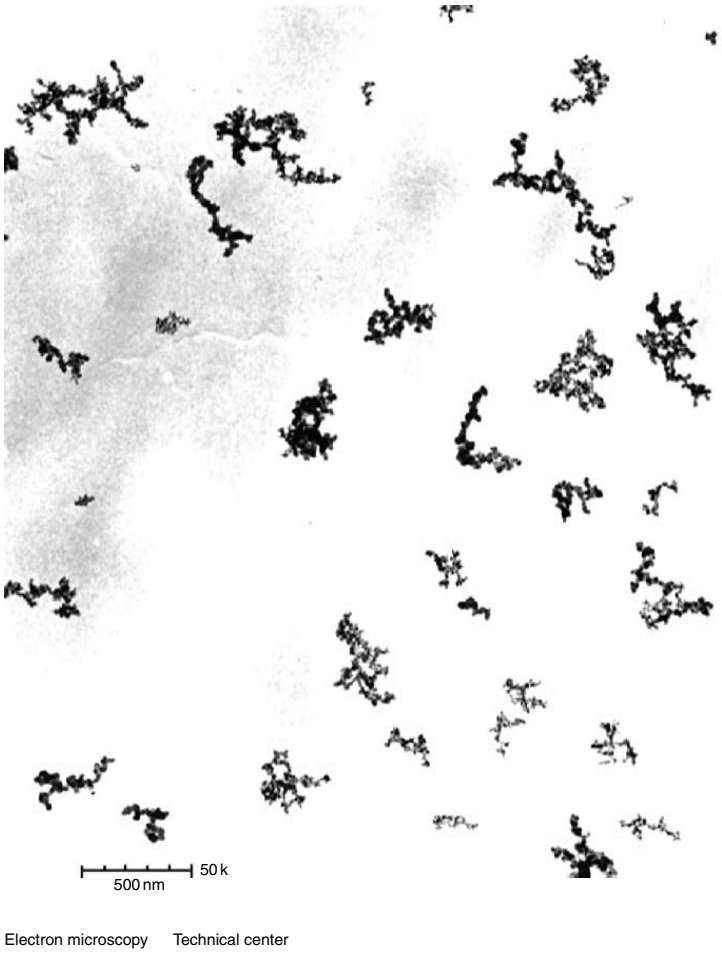


FIGURE 17.13 TEM of alumina particles.

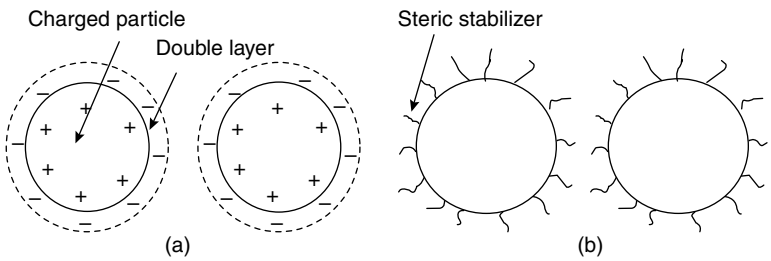


FIGURE 17.14 Stabilization of particles: (a) Electrostatic; particles are separated by repulsion of charged double layers. (b) Steric; particles are stabilized by attaching a molecule to the surface. The molecular layer causes an entropic repulsion as the particles approach each other.

TABLE 17.3 Typical Compositions for Chemical–Mechanical Planarization or Chemical–Mechanical Polishing (CMP) Slurries for Different Applications

CMP Applications	Common Abrasives	Abrasive Size (median, nm)	Percentage Solids (wt.)	pH (point of use)	Typical Oxidizers
ILD/STI	Silica	40–400	5–35	9–11	
	Ceria	20–1000			
Polysilicon	Silica	40–400	1–20	8–11	
Tungsten	Alumina	50–700	1–7	1–5	$Fe(NO_3)_3$
	Silica	40–400	1–20	1–5	KIO_3 H_20_2 K_3FeCN_6
Copper	Alumina	50–700	1–7	3–9	H_20_2 $Fe(NO_3)_3$ KIO_3

In addition, complexing agents may be used to facilitate the removal of metal from surfaces, especially in the case of copper. Corrosion inhibitors like benzotriazole (BTA) are also commonly used for copper CMP [50–52]. Additives may also be used to change selectivity to other films [53].

Choosing and balancing all these additives makes slurry formulation as much art as a science today, although there have been significant efforts in the recent past to bring more of a systematic and scientific approach to slurry formulation. Developing good slurry requires balancing the combinations of these and other additives so that they give the required performance as well as the requisite physical and chemical characteristics. Table 17.3 shows a summary of the typical compositions for various slurries for different applications.

17.2.3.2 Manufacturing Issues Relating to CMP Slurries

As previously stated, the nature and composition of slurry often has a significant effect on manufacturability of a CMP process. This section examines some of the important factors to be evaluated with respect to CMP slurry.

17.2.3.2.1 Slurry Distribution and Monitoring

In a large scale manufacturing operation, the slurry is delivered to the polisher by means of a slurry distribution system (SDS). Today's distribution systems are sophisticated pieces of equipment, far more advanced than other bulk chemical delivery systems [54]. Figure 17.15 shows a typical SDS for a factory.

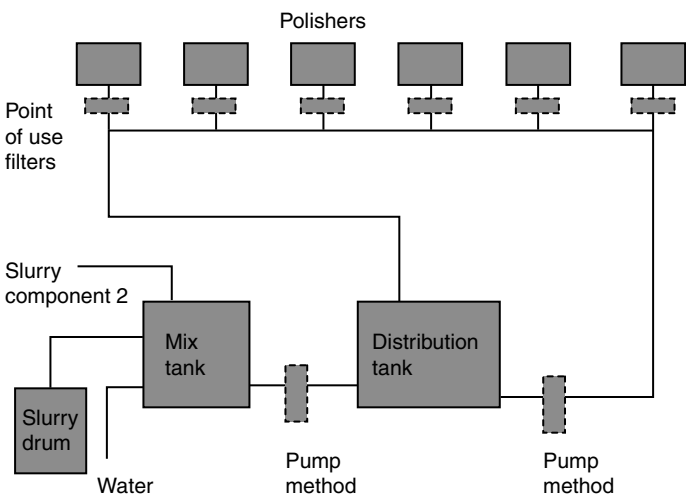


FIGURE 17.15 Layout of a slurry distribution system (SDS) in a typical fab.

All the components of the slurry (or water, if dilution is needed) are mixed in the mix tank and pumped over to the distribution tank. A third "day tank" may be used after the distribution tank for very high volume manufacturing to keep up with volume requirements; this directly feeds the polishers. POU filters are used at the polisher to remove any large particles that may have been generated in the system, thus helping to reduce defectivity. Slurry is continuously recirculated through the distribution loop to keep it suspended.

Handling of slurry in a SDS is a delicate task and care must be taken to ensure that the slurry is not "damaged." This means not subjecting it to shear that is too high (may cause aggregation) or too low (may cause settling), keeping the headspace in the tanks moist and performing regular maintenance on the system. Typical maintenance procedures include clean out and flushing of the SDS as well as polisher. These steps can significantly reduce defectivity in the CMP process. Slurry re-circulation in the SDS may be implemented by means of either double diaphragm or bellows pumps, as well as pressure/pressure or vacuum/pressure methods. Gear, vane, or centrifugal pumps generate high shear that can cause aggregation in the system. Typically, it is recommended that a minimum flow rate of 1 m/s be used for slurry re-circulation. Most manufacturers of SDS also recommend a system with about twice the maximum required slurry flow to reduce variation in dispense rates at the tool [55].

In addition to distributing slurry to a polisher, most SDS also incorporate added functionality that allows monitoring and maintaining the slurry quality; this is important for maximizing device yields [40]. The most commonly measured online parameters in a SDS include pH, specific gravity, and particle size distribution and oxidizer concentration. ZP of the slurry may also be measured as an indicator of abrasive quality and ionic contamination. Various papers have been presented on monitoring slurry "health" of different oxide and metal slurries [56,57]. They show that pumps, valves, fittings, and other components of the system have a strong effect on slurry properties. Optimization of the system is needed to reduce aggregation and minimize process variations.

Detailed recommendations regarding design and maintenance of SDS, piping, mixing, and dilution procedures, monitoring of oxidizer concentrations, filtration of the slurry, filter change-outs, and re-circulation of slurry are available from either slurry manufacturers and/or SDS manufacturers. For the user of a CMP system, it is critical to view the SDS as an integral part of the CMP manufacturing process and treat it with as much care as a polisher is. This can greatly minimize process variations, ensure smooth functioning of manufacturing, and reduce unscheduled downtime.

17.2.3.2.2 Slurry Parameter Measurements

To accurately monitor the behavior of slurry, it is important to measure both its physical and chemical properties. Measurement of physical properties (as compared to the chemical properties) has received a greater attention recently for two reasons: (i) these properties are easier to measure and (ii) there are fewer confidential information restrictions involved in divulging the physical properties. Most slurry certificates of analysis provided to end users today, list a greater number of physical properties.

The most easily measured physical properties are pH, weight per gallon, specific gravity, conductivity, and percent solids. Viscosity is often measured by using a Brookfield viscometer, which is a good tool given that most slurries behave in a Newtonian manner (viscosity is independent of shear rate). For non-Newtonian slurries, however, viscosity needs to be measured as a function of shear rate using a shear rate or a stress controlled rheometer. In such cases, it is incorrect to quote a number for viscosity without quoting the shear rate (Figure 17.16). It is important to characterize the viscosity of the slurry under the conditions of shear it is subjected to in the distribution loop as well as on the polishing table. The viscosity is important in determining slurry transport on the pad and could even affect planarization [58]. The effect of slurry viscosity modification on the removal rate has been quantified for both W and oxide CMP [59]. The effect of modifying the viscosity may be greater on metal CMP slurries than on oxide CMP slurries due to the larger chemical effects that are built into metal CMP.

In recent times, the measurement of particle size has received a great attention both from end users and suppliers of CMP slurries. Despite various studies [42], a vigorous debate exists on the effect of particle size on CMP. A large variety of techniques exist to measure particle size distributions for slurries: to

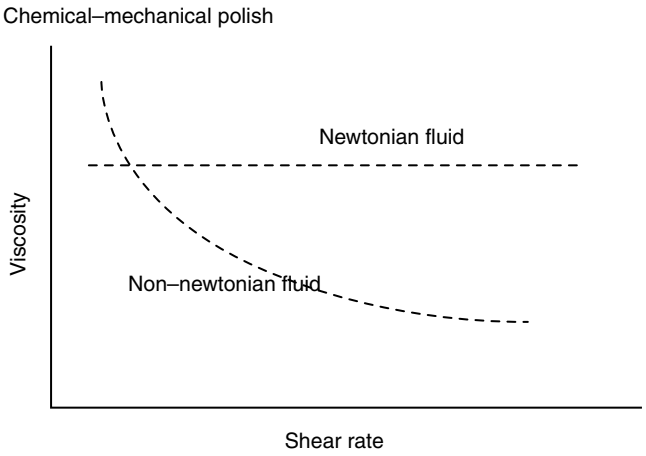


FIGURE 17.16 Viscosity vs. shear rate profiles for Newtonian and non-Newtonian fluids.

mention a few these include laser diffraction, dynamic light scattering, hydrodynamic fractionation, sedimentation, and acoustic methods. The estimation of particle size depends on the measurement technique used [60]. Each of these techniques offers a set of advantages and disadvantages; often more than one technique is needed to completely characterize the behavior of the abrasive in dispersions.

All the techniques listed above measure a signal and de-convolute this to fit a statistical distribution (often log normal or bimodal) to the slurry dispersion. While this is effective for estimating mean particle size, it does not allow measurement of large particles in the "tail" of the distribution (Figure 17.17), because they are too few to be captured by a statistical distribution. In order to effectively measure this part of the distribution, a complementary "counting" technique must be used [61,62]. Although both end users and slurry suppliers have used these techniques, there is little agreement among users correlating the number of large particles to defectivity. General trends have shown that the removal rate tends to increase with particle size up to a certain point, after which it decreases.

17.2.3.2.3 Environmental, Safety, and Handling Issues

As CMP has moved into large scale manufacturing at more and more factories across the world, waste disposal of spent slurry, by-products, and metals (from the wafer surface) have become important issues. Some fabs estimate that CMP could account for as much as 30%–40% of the total water used for IC

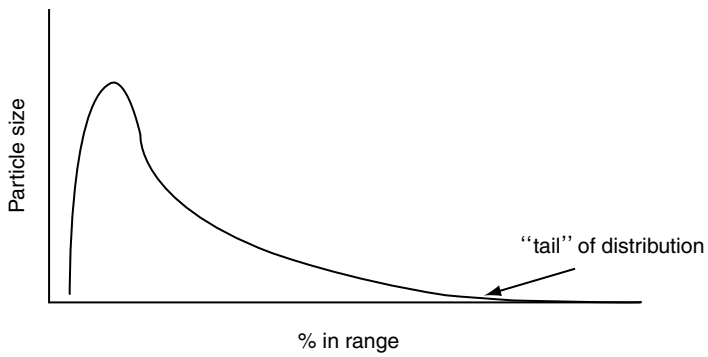


FIGURE 17.17 Typical particle size distribution of a CMP slurry.

manufacturing [63]. The issue of purification and disposal or recycling of this water becomes larger as new materials like copper, gold, and tantalum are adopted in IC manufacturing. The most common limits for waste streams are those for total suspended solids (TSS), heavy metals (like copper, chromium, zinc, and nickel), pH, and total organic content.

Environmental Health and Safety regulations differ in every city and country, and waste systems must be designed to meet these regulations before any effluents can be discharged. All materials used must be listed on the EPA Toxic Control Substance Act (TSCA) inventory list. Similar lists exist for the European Union, Korea, Japan, and Canada.

The most common methods for removing suspended solids include flocculation as well as ultra-filtration and micro-filtration. Typically, flocculation involves adding a flocculent and adjusting the pH to the level where the flocculent works best. It may then take several hours for the abrasive in the slurry to settle out. At most sites, the limit for TSS is <500 mg/l of effluent, although it may be much more stringent in certain locations. (San Jose, for example, has a limit of 0 mg/l TSS.)

After the solids are separated, the next step in water purification normally involves ion removal. Either ion exchange or reverse osmosis may do this. The advent of copper CMP creates a particularly acute problem since the waste stream contains large amounts of copper. Motorola [64] reports removal of this copper waste by a process called electro-winnowing, which reduces and plates the copper at the cathode of an electric cell. There is a possibility that the copper reclaimed in this manner can be reused.

Some companies are also looking at the possibility of reusing water from a polisher [64]. While the polisher is in idle or pad wet mode, there is a continuous stream of DI water that is used to keep all parts of the polisher wet and reduce any potential dry particles. This water has been found to have high purity and offers the potential of recycle and reuse, thus reducing waste treatment costs for CMP.

Slurry recycling has also been investigated as a potential means of both reducing waste treatment as well as reducing costs. The most commonly discussed method involves more than one use of the slurry, with various degrees of treatment being involved between each pass. This treatment could involve one or more of the following processes: filtration, oxidizer replenishment, removal of metals and replenishment of buffers, and other necessary additives. End users have also discussed the concept of leasing the slurry from a manufacturer; using it and then sending it back for regeneration.

The biggest hurdle to increase slurry recycle lies in ownership of the process. Given the finite probability of yield loss, users are reluctant to adopt an idea as radical as slurry recycle without assurance from a supplier of its viability. Suppliers, on the other hand, are reluctant to provide any assurance given that they have little control over the recycle and polishing process. The technology for slurry recycle is today in its infancy and it will require close cooperation and development between an end-user and a supplier to achieve success.

17.2.3.2.4 Manufacturability Issues Related to Slurry

The factors necessary for slurries to succeed in a large scale IC manufacturing are often very different from those required to demonstrate performance on an R&D level. While many formulations may perform adequately on a small-scale evaluation at the R&D stage, not all of them may actually be able to consistently perform in a high volume IC manufacturing.

The most important aspect of a manufacturable CMP process is the ability of the supplier to consistently deliver slurry that has the same performance. This means that the slurry must perform the same from lot-to-lot with no surprises for the end user (Figure 17.18). This reduces the need for monitor wafers and greatly lowers the cost of CMP. One of the important factors that determine this consistency is the sensitivity of the slurry properties to its individual components; this is determined by a formulation window study. Robust slurry is formulated so that it is in a relatively "flat" region of the sensitivity curve.

In addition to the sensitivity to the formulation described above, certain slurries are inherently more sensitive to the process conditions (for example, the polisher settings or the pad). This may occur due to the inherent instability of the slurry or to the nature of the formulation itself. If this occurs, it becomes difficult to maintain a consistent process. This effect is often neglected in the choice of slurries, but plays a large role in minimizing time spent troubleshooting problems in manufacturing.

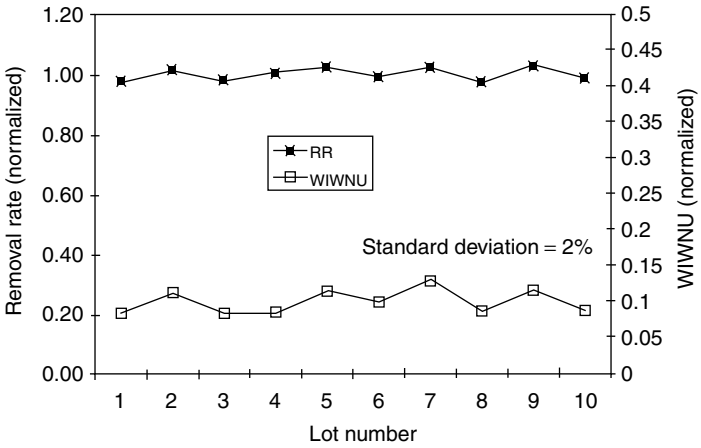


FIGURE 17.18 Lot-to-lot consistency of a CMP slurry. The plot shows 10 lots of slurry manufactured at different times and polished on the same pad at the same time.

The number of components in the slurry affect the simplicity of the CMP manufacturing process. Two (or more) components bring an element of complexity to the CMP process; the performance may be affected by the mix ratio of the two components. Specifically, the robustness of the process depends on its invariance to this mix ratio. Using of single component slurries reduces this risk. In addition, all two component slurries have a limited pot life (useful life of slurry after the components are mixed). The shorter the pot life, the greater the demands placed on the end user of the slurry. Some slurry, notably those that use hydrogen peroxide as an oxidizer, can have their pot life extended by adding more hydrogen peroxide to the mixed slurry. Concentrates are a special case of two component slurries where the second component is water. While diluting a concentrate, effects like order of mixing and pH shock become important in minimizing agglomeration of the slurry.

Colloidal stability of the slurry is another important factor in reducing the overall cost of ownership of a CMP process. While this may not be explicitly factored into a CoO calculation, it strongly affects down time and maintenance cost of SDS and polishers. The slurry that is colloidal and unstable tends to have more particles settling out in dead legs of a SDS. Unstable slurries often need more frequent changes of filters and may need special mixing equipment.

The shelf life of slurry determines how long the slurry can be used after manufacture. Some slurries need to be maintained within a given temperature range during shipping and storage to maintain their performance. This adds to shipping costs and makes it logistically more difficult to use in a manufacturing environment. The shorter the shelf life of the slurry, the narrower the window that it can be used; in this case, inventory control becomes important, especially if the slurry has to be shipped around the world. An ideal shelf life for all slurries is at least 1 year.

Given that CMP slurries contain many chemical additives, safety handling and environmental issues become important. They could have a significant impact on the CoO of a process; with a corrosive chemical, for example, different components of the system could be affected over time, requiring more frequent hardware replacement and higher downtime. Certain corrosive chemistries even affect platens on polishers by corroding them over long periods of exposure. To conclude, while the polishing performance of the slurry is important in determining the CoO of a CMP step, many of the other factors discussed here can have a large effect on final cost. Using a slurry that is easy to implement in manufacturing may lead to higher uptime, lower problems, and eventually a smaller CoO even though the removal rates may be lower and the actual polish times per wafer longer. These manufacturability factors are often overlooked during evaluations, but become more important, as the process moves to high volume manufacturing.

17.3 Mechanisms and Models

In the introduction it was pointed out that CMP has become an established process technology for fabrication of integrated circuits (IC). It is a vital technology which enables (STI) without the characteristic "birds beak" topography, global planarization of pre-metal dielectrics (PMD), ILD, and metal inlaid (i.e., damascene) interconnects. The state-of-the-art polishing tools were described in the second section, along with the status of supply of expendable materials under the title, Equipment and Consumables. In that section the 15-year experience the SC industry has had with CMP was presented in a pragmatic way. Recommended practices were based on the open literature and the interpretation and experiences of the authors. In this section it is our goal to provide the process manager and engineer with a framework with which to understand the "theory" of SC CMP.

We will attempt to provide a consistent theoretic basis for CMP of microelectronic materials in this section. The term theory is used very broadly. The goal of this section is for the reader to come away with the thought that process results are predictable on the basis of some fundamental chemical and mechanical properties of the wafer, the pad, the slurry, and tool design; rather than on the basis of some empirically generated rules-of-thumb.

This section is organized into three major headings: "The fundamentals of polishing wafers," "Polishing patterned wafers," i.e., wafers with topography, and "Chemical effects in polishing." This framework is historical. All that the microelectronic scientist and engineer know about polishing is derived from the early work on polishing and shaping flat blanks of glass into optical lens. The Preston equation is an accurate depiction of the mechanical nature of the optical glass industry [14]. Using a Hertzian penetration model several authors have come up with physical models which lead to derivations equivalent to the Preston equation or a power term modification of it. Understanding the impact of soft and viscous-elastic polyurethane pads on polishing non-patterned or blanket wafers is our first step; we then move on to discuss the impact of the pad on patterned wafer polishing. This pad model, as applied by recent authors to patterned wafers, is next discussed. The section on mechanical modeling of patterned wafers concludes with a discussion on the success of statistical methods to de-convolve step height change and non-uniformity into wafer scale and die level pattern effects. The usefulness of the concept of planarization length in characterizing flexible and compressible pads is also discussed. Using this sound foundation of mechanical principles we move on to the consideration of chemical effects. Following the historical sequence, the chemical effects are discussed in the material order: oxide dielectrics, tungsten, and copper.

17.3.1 Polishing Non-Patterned Wafers

17.3.1.1 Mechanics of Polishing

Thin, silicon oxide films, grown at 1100°C in a tube furnace, on flat silicon wafers, have mechanical and chemical properties very much like quartz. Chemical vapor deposited silicon oxide films, grown at 350°C–450°C in plasma from tetraethyloxysilicon (TEOS) and ozone precursors, also has quartz-like chemical and mechanical properties, only the deposited film is somewhat less dense and softer. Thus, it is reasonable to apply the same models that have been used to understand polishing quartz and optical glass lens to polishing silicon oxide on flat silicon wafers. The first principle for glass polishing, namely the rate of material removal, is described by the Preston equation [14], introduced earlier as Equation 17.2 and written here as:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = K_{\mathrm{p}} P \frac{\mathrm{d}s}{\mathrm{d}t} \tag{17.4}$$

where T denotes the thickness of the material, P is the polishing pressure, s is the distance traveled, and t denotes time. In general, the removal rate of a material is proportional to the pressure and the relative velocity between the pad and the wafer, $V = \mathrm{d}s/\mathrm{d}t$. Any physical and chemical considerations are simply put into the Preston constant K_{p}, which is the part of the removal rate dependence that is assumed to be

independent of pressure and velocity. In practice, the Preston constant contains the influence of chemical effects, hardness of pad, size of abrasive particle, slurry chemistry, film hardness and other process, consumable, and tool parameters.

A criticism of the Preston equation is that it was formulated empirically rather than from first principles, i.e., mathematically derived from a physical model of the polishing process. Several authors have proposed equations from first principles that predict the dependency of the polish rate on pressure and velocity. Brown et al. [65,66] developed a model in which spherical particles of a Gaussian size distribution are fixed in a viscous Newtonian fluid (pitch) under a uniform load. The particles move across the work piece surface at velocity ds/dt. At some pressure all the particles become load bearing and the pitch to work piece surface gap is fixed by the diameter of the smallest particles. It was shown that the removal rate is given by:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \left(\frac{P}{2E}\right)\frac{\mathrm{d}s}{\mathrm{d}t} \tag{17.5}$$

where E is the Young's modulus of the work piece. Comparing Equation 17.4 and Equation 17.5, we see that Preston's constant is equivalent to the inverse of twice Young's modulus. Brown et al. [66] developed this model first for metal polishing; however, Brown and Cook [67] proposed that the model also applied to glass.

Mechanical effects have also been shown to play an important role in controlling microelectronic oxide polishing. Liu et al. [68] illustrated the linear relationship between surface hardness measured by nano-indentation and the removal rates of the non-doped dielectric films under a well-controlled CMP process. It is also widely reported that the doped oxides, such as PSG and borophosphate silicate glass, have higher polish rates than the non-doped oxide. Figure 17.19 shows the polish rate of PSG as a function of phosphorous concentration. Since both P_2O_3 and P_2O_5 are softer glasses than SiO_2, the incorporation of P_2O_3 and P_2O_5 into the SiO_2 network softens the film [1]. In order to correlate the mechanical hardness with the polishing characteristics, a mechanical model based on a statistical method and the elastic theory to describe the wear mechanism during polishing was proposed by Liu et al. [69]. Kallingal et al. [70] studied the substrate effect on the hardness and polish rate of oxide films. Their experimental results show that the substrate has no influence on the polish rate, although it strongly

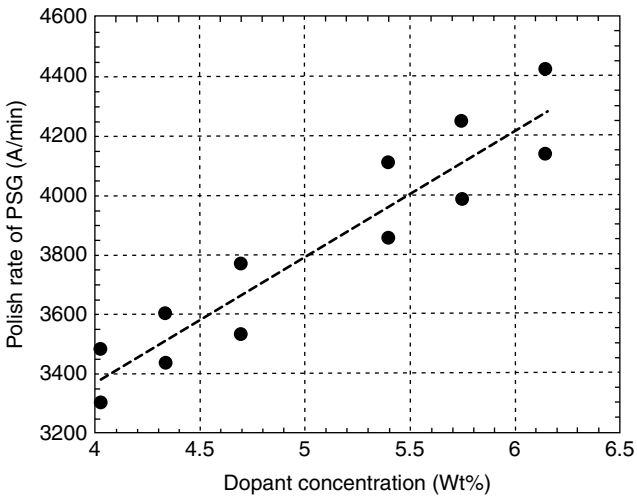


FIGURE 17.19 Polish rate of PSG as a function of phosphorus concentration: pressure; 4 psi, table speed; 40 rpm, carrier speed; 63 rpm.

influences the hardness of the film being polished. The observations suggest that the depth of deformation of films by abrasive particles during polishing is much shallower than that by the diamond tip used to make micro hardness measurements. The intrinsic hardness of a film, independent of substrate materials, is obtained when the indentation depth is less than one-tenth of the film thickness. A greater influence of substrate hardness on polish rate is expected for oxide films on softer substrate, such as aluminum, where greater deformation of the film by abrasive particles would take place.

Runnels and Eyman [71] proposed a model based on wear and erosion, i.e., a tribology model. Their starting point for removal rate, RR, is some function of the normal stress σ_n and the shear stress τ, with the removal rate being a simple linear product of these two stress tensors, with proportionality constant K_t.

$$\mathrm{RR} = K_t \sigma_n \tau \tag{17.6}$$

They used a finite element methodology to compute the slurry film thickness as a function of the load and the angle of attack for a gimbaled carrier holding a wafer with fixed radius of curvature. For the baseline pressure, 7 psi, zero moment between the carrier and the table, and a velocity of 20 rpm with a wafer radius of curvature of 500 m, they find that the gap is 63 μm with an angle of attack of 0.01 degrees. Based on this model, particles smaller than 63 μm move freely about in the gap and are not fixed as are the Hertzian indenters of the Brown and Cook model. The authors admit that the model lacks accuracy and should be valid only in predicting relative effect on tool parameters.

Tseng et al. [72] combined the tribology model of Runnels and Eyman with the traveling indenter model of Brown and Cook. Their result gives a sub-linear dependency for both pressure and velocity:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = MP^{5/6}\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^{1/2} \tag{17.7}$$

In this model, removal rate seems to have a weaker dependence on the speed. Most importantly, they showed that the proportionality constant M is a function of both speed and pressure. M increases as the pressure increases for thermal oxide, where the increase of M is a stress-assisted chemical effect. Details of the chemical effect are, however, not fully discussed in this model.

Another model, based on the fixed particle indenter, was proposed by Shi et al. [73,74] and Zhao and Shi [75] that includes the effect of pad and wafer hardness into the Preston equation,

$$\mathrm{RR} = K\frac{(E_p P^2)^{1/3}}{E_w} V \tag{17.8}$$

where E_w and E_p are the elastic modulus of wafer and pad, respectively. This is the first work in the literature trying to incorporate the mechanical properties of the pad into the Preston constant. Thus, it is a reasoned extension of the Brown and Cook model. In contrast to the conventional Preston equation, the pressure dependence of the softer pads of microelectronic polishing is found to be proportional to $P^{2/3}$. It was concluded that the linear dependence of removal rate on pressure in Preston's equation is only applicable to polishing with a pad whose hardness is similar to or harder than that of the abrasive particles and the polished surface. The derivation in [75] also proposes a threshold pressure P_{th} in the removal rate vs. pressure equation,

$$\mathrm{RR} = K_e V(P^{2/3} - P_{th})^{2/3} \tag{17.9}$$

where

$$K_e = K\frac{E_p^{1/3}}{E_w} \tag{17.10}$$

Zhao claims that this threshold pressure is the minimum pressure for the average particle to become fixed in the soft visco-elastic pad and create the Hertzian indentation in the wafer surface. Experimental

data from the literature is potentially consistent with the two-thirds power dependence on pressure, with positive intercepts at zero removal rates [76–78].

If Shi et al. and Tseng et al. could substitute the Hertzian indenter model for the normal stress tensor, and combine it with the shear tensor, our technology would have a "first principles" derived rate removal equation. The relationship should show the $P^{2/3}$ dependence, rather than the $P^{5/6}$ dependence, with the $V^{1/2}$ dependence for typical microelectronic CMP pads, i.e.,

$$\mathrm{RR} = C_e P^{2/3} V^{1/2} \tag{17.11}$$

where C_e is a new proportionality constant which contains the Young's moduli of the pad and the film on the wafer being polished, and includes all the chemical effects. However, experimental evidence for such a non-Prestonian pressure and velocity dependence is lacking.

17.3.1.2 Modeling of CMP Edge Effect

Ideally, a non-patterned wafer should be polished homogeneously and uniformly across the entire wafer—that is, the removal rate should be identical on every measurement spot within a wafer. In reality, CMP tends to degrade the within wafer non-uniformity (WIWNU) at the edge of the wafer unless there is a very non-uniform thickness before CMP. The edge effect describes the local thickness variation at the very edge of the wafer. Several workers have sought to develop mechanical models to explain this well-known CMP phenomenon [79,80]. Figure 17.20 shows two typical thickness profiles of wafers polished by two different polishers. It is clear that most of the thickness variation occurs at the very edge of the wafer. Wang et al. [80] attribute this phenomenon to the pad deformation profile and pressure distribution, or Van Mises stress, at the very edge of the wafer (Figure 17.21). The edge effect is generic to most commercial polishers, but it can be minimized if we understand the pad deformation profile and pressure distribution at the edge of the wafer. As discussed in Section 17.2, this effect can be greatly reduced by optimizing the pressure distribution at the edge of the wafer through the use of carriers with wide pressure-bearing retaining rings [81]. Currently, most commercial polishers offer decent performance on WIWNU at a 5–10 mm edge exclusion. In order to place more yielding dies on a wafer, we will need to improve the edge profile in the future (i.e., with edge exclusion <5 mm). This requirement puts a stringent demand on the development of CMP process and polishing carrier design.

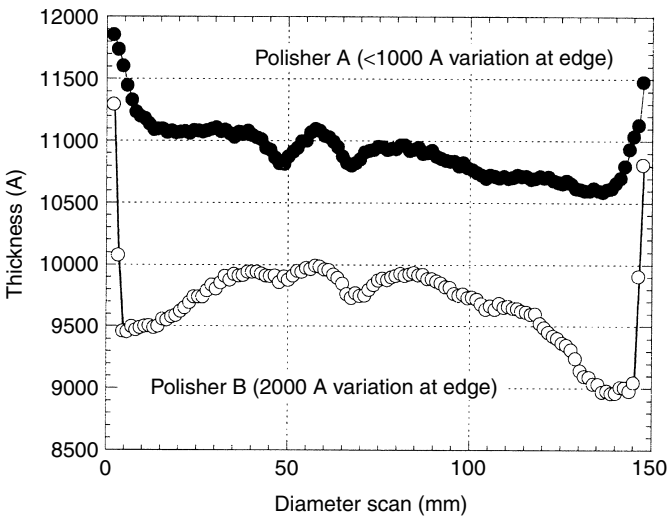


FIGURE 17.20 Diameter scans of wafers polished using two different style carriers. A is a bladder-backed, contact retaining ring carrier. B is a film-backed, non-contact retaining ring carrier.

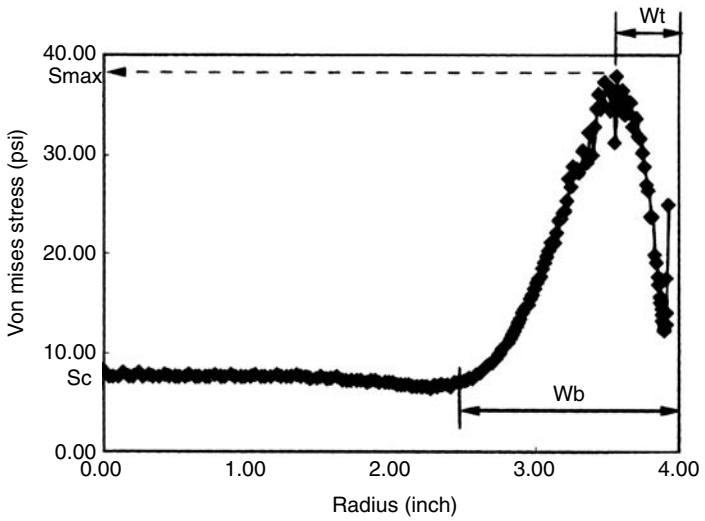


FIGURE 17.21 Distribution of pressure profile at wafer edge. (From Wang, J. L., Holland, K., Bibby, T., Beaudoin, S., Cale, T., *J. Electrochem. Soc.*, 144, 1121–7, 1997.)

To summarize, for non-patterned wafer polishing:

1. The Preston equation is a good model for mechanical lapping or polishing on inflexible platens with hardness comparable to the abrasive particles, both of which are harder than the surface to be polished.
2. The removal rate of microelectronic materials by CMP, using visco-elastic pads, is non-linear in the load applied to the wafer and in the velocity of the platen.
3. The removal rate for microelectronic material with compressible pads, $E_p \leq 400$ MPa, i.e., IC-1400 or IC-1000/Suba IV pad, have been proposed to be dependent on the 2/3-power of pressure and 1/2-power of platen velocity. However, there is little experimental evidence for statistically significant deviations from the linear dependence of removal rate on pressure and velocity.
4. The alternate bands of higher followed by lower than average removal rates at the wafer edge has been modeled as Van Mises stress. It is correctable by using carriers with wide retaining rings that pre-compress the pad.

17.3.2 Polishing Patterned Wafers

In the previous section, we only discussed the polishing models for non-patterned wafers. Recent efforts on CMP modeling focus on the prediction of thickness removal and step height variability for patterned wafers. In practice, CMP suffers from the degradation of within die thickness variation or non-uniformity (WIDNU), due to the variation of pattern density. Understanding the effect of pattern density on polishing characteristics can help us to have a wider process margin on etch and lithography, leading to improvements on device performance and yield.

17.3.2.1 Models due to Pad Mechanics

Burke [82] was the first to publish a model for the pattern-dependent CMP polish rate. He proposed that the local polish rate of the patterned wafer depends on the "degree of non-planarity," which is the ratio of down area to total wafer area. This model takes into account the type of the non-planarity of the wafer surface and adjusts the polishing rate accordingly. The model is empirical and does not address the dependence of polish rate on the pressure.

The first quantitative model trying to predict both relative and absolute polish rates of arrays of features with different sizes and pattern densities was proposed by Warnock [83]. It is a 2D microscopic mathematical model. This model adopts the notion that planarity is a function of pad compression to predict planarization. Even though this model is completely phenomenological, it illustrates the importance of pattern density and feature size in CMP.

Runnels and Eyman [71] and Runnels [84] proposed several models to account for the stress in the polishing pad and the fluid flow as well as the removal of material by erosion. Runnels et al. tried to include physical effects, such as fluid flow and pad deformation in the CMP modeling. They were able to predict the experimental data of Warnock. Based on these models, a physically based erosion simulator was developed by Runnels et al. [85].

Another approach in CMP modeling deals with the effect of pad asperities on polish rate, Yu et al. [86]. Although this is the first model that considers the effect of polishing pad roughness and the dynamic interaction between pad and wafer, the discussion on the slurry transport and pad porosity has not been provided in this model. The most important contribution of this model is to point out that the pad deformation can be a key to SHR. They reported that polishing with high-pressure results in a high degree of pad deformation and leads to the high planarization rate (i.e., SHR rate).

Both models proposed by Runnels and Eyman [71] and Yu et al. [86] are physically based, but these models only deal with local polishing characteristics. Global planarization is the main advantage of CMP. Models that only account for chemical reaction, pad asperity, or local abrasive action cannot explain the planarization over long (mm) lateral distances; explanation of global planarization is possible only if some non-local effects are involved. Warnock's model provides non-local information, but the model is phenomenological and the fitting parameters cannot relate to real parameters in the CMP systems (wafer–slurry–pad).

Chekina et al. [87] proposed a model where the pad elastic deformation and the wafer surface evolution are considered based on contact mechanics and the theory of wear-contact. In their work, they successfully simulated profiles of the erosion over a wide field and the recess within a small geometry. Compared with Warnock's model, this model considers the interactions among wafer, slurry, and pad. However, this paper only semi-quantitatively illustrates the known CMP effect. The details about the fitting parameters are not provided in the paper.

Another approach in CMP modeling is to predict the time evolution of step height. Tseng et al. [88] proposed a mathematical derivation to predict the evolution of step height and thickness during polishing. This model is based on the assumption that the polishing pad is completely compressed and conforms to the pattern; that is, the pad touches both up and down area of the topography simultaneously. Unlike the previous model proposed by Yu et al. this model does not consider the microscopic effect, such as the pad roughness. This model only considers the deformation of pad from a macroscopic point of view. Most importantly, Tseng et al. provided an analytical formula to express the evolution of step height. The work of Tseng et al. provided an excellent agreement between experimental data and model prediction. However, there are two major drawbacks for this model: (1) it does not fully account for the pattern density effect and (2) it fails if the pad is not completely conformal (i.e., pad does not touch the down area of the topography during the entire polishing step). In this model, Tseng et al. only use the "cell" and its periphery to account for the areas of two different pattern densities in the DRAM. Since the DRAM typically has more uniform pattern density than the application specific integrated circuit or some logic applications, the assumption about the conformance of pad may be applicable to the particular measurement site for the memory applications. Due to the lack of a rigorous definition of pattern density, this model is not general enough to account for a broad range of CMP effects.

Grillaert et al. [89] at the Interuniversity Microelectronics Center (IMEC) showed that the time evolution of step height depends on the polishing characteristics of the pad. Unlike the model of Tseng et al. they discussed the polishing characteristics of both "compressible" and "incompressible" pads. For the incompressible (inflexible) pad, the SHR is linear with polishing time. For the compressible (conformal) pad, the step height decays exponentially which has been previously shown in the work of Tseng et al.

The mathematical expression of this model is shown as:

$$h(t) = h_0 - \frac{r}{a}t \quad \text{for incompressible (inflexible) pad, } h > h_t \tag{17.12}$$

$$h(t) = h_0 e^{-t/\tau} \quad \text{for compressible (conformal) pad;} \quad h \leq h_t \tag{17.13}$$

$$\tau = lP/Er \tag{17.14}$$

$$h_t = lP/aE \tag{17.15}$$

where h is the step height as the function of time; h_0 is the initial step height before polishing; r is the polish rate of blanket wafer; a is the layout pattern density; P is the polishing pressure; and E is the elastic modulus of pad. The time constant (τ) can be predicted from Equation 17.14. The value h_t is the step height when the contact of down area occurs; that is, the height at which the transition between linear and exponential region occurs. As pointed out in Section 17.2, real pads are neither totally conformal nor inflexible; rather they are flexible with varying degrees of compressibility, which is expressed by Young's elastic modulus. Before the pad touches the down area, the polishing pad can be modeled as an inflexible pad. When the polishing pad touches the down area, the polishing pad should be modeled as a conformal pad.

Compared to the model proposed by Tseng et al. the pattern density is a well-defined parameter in the mathematical expression of the model proposed by Grillaert et al. Most importantly, Grillaert et al. have attempted to predict the global CMP effect, such as the within die non-uniformity (WIDNU), which is defined as:

$$\text{WIDNU} = h_0(b-a) \tag{17.16}$$

where a and b are the pattern densities of two different areas. Equation 17.16 shows that the WIDNU can be predicted if the pattern densities (i.e., a and b) are known. Also, the WIDNU should be a constant after the pad is conformal to all wafer areas, i.e., "planarization" is reached.

17.3.2.2 Modeling Pattern Density Effects by Spatial Averaging

Among the recent studies of CMP modeling, a semi-empirical approach from Boning et al. at MIT has shown good comparisons between theory and experiment. In the early work, the group both proposed a physically based model to predict the thickness evolution as a function of pattern density and also provided a methodology to experimentally characterize these CMP pattern dependent effects. First, Stine et al. [90] developed a methodology to analyze the spatial variation within patterned (and not just blanket) wafers. By separating the wafer-level thickness variation from the die-level variations, the effects due to various device layouts can be identified. Various masks for ILD CMP are proposed to characterize the effect of area, pattern density, pitch, and aspect ratio (i.e., perimeter/area) on polishing characteristics [91–94]. Based on experiments, pattern density is found to be a strong dominant factor, while structure area, pitch, and aspect ratio play only a minor role [92]. In order to model the pattern density effect, the MIT model describes the polishing process as:

$$z = z_0 - \left(\frac{Kt}{\rho_0(x,y)}\right) \tag{17.17}$$

$$z = z_0 - z_1 - Kt + \rho_0(x,y)z_1 \tag{17.18}$$

$$\rho(x,y,z) = \rho_0(x,y) \quad z > z_0 - z_1 \tag{17.19}$$

$$\rho(x,y,z) = 1 \quad z > z_0 - z_1 \tag{17.20}$$

where K is the blanket polish rate; $\rho(x,y,z)$ is the effective pattern density that depends on the position (x,y) on the die; t is the polish time; z_1 is the initial step height; and z_0 is the oxide thickness on top of metal interconnect (Figure 17.8). This model divides the polishing into two regimes. It assumes a locally stiff pad, i.e., no contact is made to down areas between features, so that all of the pressure is born by the raised area above patterned features. First, a linear reduction of step height occurs while the local step height has not yet been fully planarized (i.e., $z > z_0 - z_1$), at a rate that is inversely related to the effective pattern density. Second, once local planarization has been achieved, both "up" and "down" areas are simply removed at the blanket polish rate. The integrated result over the entire polishing time t (assuming the chip has been

polished through to local planarity everywhere) is a final oxide thickness that depends linearly on the initial pattern density $\rho_0(x,y)$. The determination of the pattern density $\rho_0(x,y)$ is thus crucial to the prediction of die-level thickness evolution. The $\rho_0(x,y)$ at a spatial location on the die is defined as the weighted ratio of raised to total area within an averaging region.

The MIT pattern density model defines a parameter for calculation of the effective pattern density, termed the planarization length [92]. The planarization length is related to the transition range of the deformation profile of an elastic material under distributed load (Figure 17.22). The MIT model empirically extracts the global information about the polishing characteristics (i.e., planarization length), and uses this to predict the local thickness variation across the entire chip. Intuitively, planarization length represents the range over which the polishing pad bends in response to the wafer pattern and resulting topography. A large planarization length implies that the pad is flatter or stiffer over a wider area, leading to a better global planarity.

The IMEC model assumes that there are two stages during CMP: one prior to pad contact with down areas between features, and a second stage after this contact is made. The transition from a linear SHR in the first stage, to an exponentially decaying step height in the second stage, occurs at a specified contact height. While the simple pattern density model from MIT is able to capture final within-chip thickness variations, it assumes essentially a contact height of zero, and thus does not describe the exponential time-dependence of the SHR. However, the linkage to effective pattern density in the MIT model provides spatial information about the polishing characteristics (rather than focusing on the time evolution of any single feature step height), and the planarization length defined in the MIT model is a good indicator to extract and summarize the global chip-scale variation dependence on the CMP process.

The creation and reduction of WIDNU, due to different polishing rates of regions with different pattern densities, has been of substantial concern in oxide and other CMP processes. Fang et al. [95] reported that after the local planarity has been achieved, the WIDNU decreases as the polishing proceeds. This result seems to contradict the prediction shown by Equation 17.17 through Equation 17.20. In fact, Fang et al. pointed out that the "planarity" (or step height measurement) is an ambiguous parameter. Depending on the measurement location of step height, a different conclusion can be reached about the polishing process. This is because the step height measurement can provide only local information about the step height, rather than the global thickness variation during CMP. The IMEC model primarily considers the local

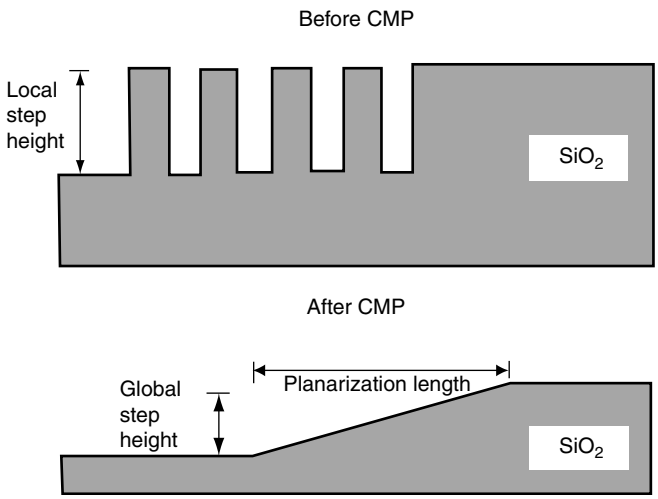


FIGURE 17.22 The definition of planarization length. Typical planarization lengths, for an elastic material under distributed load, is of the order of several millimeters. (From Ouma, D., Modeling of Chemical Mechanical Polishing for Dielectric Planarization, Ph.D. thesis, MIT, Cambridge, MA, 1998.)

information about the local step height and thickness, rather than global thickness variation. Indeed, Fang et al. illustrated that the planarization length defined in the MIT model is effective in understanding the global polishing characteristics: a CMP process with a larger planarization length results in a lower WIDNU. While capable of predicting the spatial variation within a die well, the step height dependence noted by Grillaert et al. should also be accounted for.

Smith, et al. [96] provide a more detailed mathematical expression which succeeds in combining aspects of both pattern density and step height highlighted in the early MIT and IMEC models. Thus, both spatial information and time-dependency can be captured in a single model, with corresponding thickness prediction accuracy improvements. The mathematical expressions of this model are:

$$\mathrm{AR_u} = \begin{cases} \dfrac{Kt_\mathrm{p}}{\rho}; & t_\mathrm{p} \le t_\mathrm{c} \\ \dfrac{Kt_\mathrm{c}}{\rho} + K(t_\mathrm{p} - t_\mathrm{c}) + (1-\rho)\dfrac{h_1}{\tau}(1 - \mathrm{e}^{-(t_\mathrm{p}-t_\mathrm{c})/\tau}); & t_\mathrm{p} > t_\mathrm{c} \end{cases} \tag{17.21}$$

$$\mathrm{AR_d} = \begin{cases} 0; & t_\mathrm{p} \le t_\mathrm{c} \\ K(t_\mathrm{p} - t_\mathrm{c}) - \rho\dfrac{h_1}{\tau}(1 - \mathrm{e}^{-(t_\mathrm{p}-t_\mathrm{c})/\tau}); & t_\mathrm{p} > t_\mathrm{c} \end{cases} \tag{17.22}$$

where $\mathrm{AR_u}$ and $\mathrm{AR_d}$ are the amount of removal of up and down area, respectively, K is the polish rate on the blanket wafer, ρ is the effective pattern density; t_p is the polish time; t_c is the time when contact with down area occurs, τ is defined in Equation 17.14, and $h_1 = h_0 - Kt_\mathrm{c}/\rho$. As in the original MIT density model, the effective density is calculated by varying the "window size,", or planarization length for layout pattern density averaging. For each candidate planarization length, a constrained optimization is performed in order to find K and τ, as well as t_c of each site (note that t_c is a function of effective pattern density). The final model parameters are those which minimize the total root-mean-square (RMS) error. This combined time dependent pattern density and step height model predicts the thickness evolution of both up and down areas. By subtracting the up and down area thicknesses, we can obtain the dependency of step height as a function of time and pattern density. Notice that this model predicts three polishing regimes for flexible, compressible pads. When $t_\mathrm{p} \le t_\mathrm{c}$, the area removed thicknesses are linearly dependent on time. When $t_\mathrm{p} > t_\mathrm{c}$, but not $t_\mathrm{p} \gg t_\mathrm{c}$, the area removed thicknesses are in transition, showing both a linear plus an exponential dependence on time (and the step height has a purely exponential dependence on time). Finally, when $t_\mathrm{p} \gg t_\mathrm{c}$, the area removed thicknesses are again linearly dependent on time, corresponding to further reduction at the blanket removal rate.

The modeling of patterned wafer effects for dielectric CMP can be summarized as follows:

1. The viscoelastic properties of polish pads used in microelectronic CMP can explain WIWNU and WIDNU.
2. Hard, flexible pad material, backed with a compressible material, e.g., Rohm and Haas IC-1400, can be modeled as an inflexible pad, until step height loss allows the down area of a wafer to be touched. In the large step height regime, inflexible pads have a linearly reducing step height as a function of time.
3. Soft pads that approach being completely conformal to the wafer topography, e.g., Rohm and Haas Suba 500 series and Poliytex, have an exponentially reducing step height as a function of time, resulting from local apportionment of pressure between the up areas and down areas of patterned features.
4. The MIT pattern density mask set and the corresponding spatial averaging model characterizes the "planarization length" of a CMP process.

5. Pad materials or polish tool parameters that increase the process planarization length are indicative of lower WIDNU.
6. The SHR of the polishing process can be modeled, on a single flexible, compressible pad, as having three regimes:
   - The initial linear regime, where the down areas are not contacted, $t_p \leq t_c$
   - The exponential transition regime, where some of the down areas are contacted, $t_p > t_c$
   - The final linear regime, where all the local feature step heights have been removed, $t_p \gg t_c$.

17.3.2.3 Dishing and Erosion in Damascene Structures

In contrast to oxide CMP, metal CMP is more dynamic in that the overburdened metal over the field dielectric is composed of the conductor (W, Al, or Cu) on top of a barrier (Ti, TiN, Ta, TaN, or Ta_2N) on top of the dielectric. This leaves the surface of the damascene wafer heterogeneous compared with the homogeneous and flat surface of oxide CMP, excluding STI CMP. This complicates the modeling of metal CMP [8,47,97] due to the polish rate variation of the different materials, not to mention electrochemical interaction between barrier/adhesion layers and metal layers [98]. One may take advantage of the differing removal rates to have an effective stopping layer for metal CMP. Slurries are normally manufactured to have high metal to oxide selectivity to take advantage of the stopping layer effect. Although the stopping layer is an advantage over oxide CMP, this in turn causes dishing of the metal vias and metal lines due to the higher polish rates of the metal which in turn increases the removal rate of the now high oxide areas leading to a cyclic dishing and erosion phenomena. Thus, most effort in metal CMP is to reduce the amount of over-polish by optimizing endpoint, endpoint polish conditions, non-uniformities of removal, incoming non-uniformity, barrier thickness, metal patterns, and selectivities of differing materials.

17.3.2.3.1 Dishing/Erosion

Due to the many different patterns for the inlaid metal in a typical clip layout, dishing and erosion can cause a significant amount of non-planarity within a die. This will result in higher variation in sheet resistance from isolated lines to dense lines, and integration complications when adding multiple levels of metal. Dishing is normally measured as the step height from the neighboring oxide to the center of the metal line as seen in Figure 17.23. Erosion is the thinning of the oxide due to the non-zero removal rate during over-polish of the metal. One must be specific what line width size and pattern density one is measuring when quoting dishing and erosion values, since dishing depends heavily on line width and erosion depends heavily on pattern density (Figure 17.24).

For metal CMP, an over-polish amount is required to clear residual metal and/or barrier to increase device yield, due to the inherent polish non-uniformity at the chip and wafer scale. As mentioned previously, the over-polish time is critical to both dishing and erosion. This can be understood by the following [97]. In previous sections, the CMP removal rate given by Preston's equation is dependent on the pressure and linear velocity as well as the other factors rolled into the Preston coefficient K_p. If we start by the idealized case shown in Figure 17.24, during the over-polish step, the up areas of the patterned regions will experience a greater pressure from the pad by a factor

$$P_{\text{dielectric}} = P_{\text{applied}}/(1-\rho) \tag{17.23}$$

where ρ is the patten density for the inlaid metal feature. Thus, the removal rate for field areas and patterned areas are

$$\text{RR}_{\text{field}} = KVP_{\text{applied}}(\text{where } \rho = 0) \tag{17.24}$$

$$\text{RR}_{\text{patterned}} = KVP_{\text{applied}}/(1-\rho) \tag{17.25}$$

and the step height erosion rate will be given by

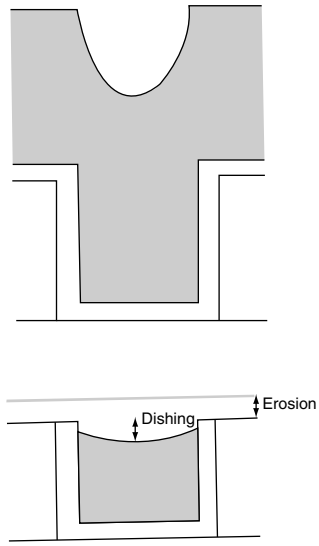


FIGURE 17.23 Definition of dishing and erosion on an isolated metal pattern.

$$\text{Erosion rate} = (\text{RR}_{\text{patterned}} - \text{RR}_{\text{field}})t_{\text{overpolish}} \tag{17.26}$$

Combining Equation 17.24 through Equation 17.26 gives

$$\text{Erosion rate} = KVP_{\text{applied}}(\rho/(1-\rho))t_{\text{overpolish}} \tag{17.27}$$

Thus one can see that erosion is linear with over-polish time and $\rho/(1-\rho)$. To minimize erosion, the over-polish time must be reduced by optimizing polish uniformity across the wafer and defining an appropriate end point, which will be discussed later. Other over-polish reduction options are: (a) increasing barrier removal rate, (b) decreasing barrier removal rate, followed by a touch polish using a slurry that has similar removal rates for barrier and dielectric, or (c) changing process conditions during

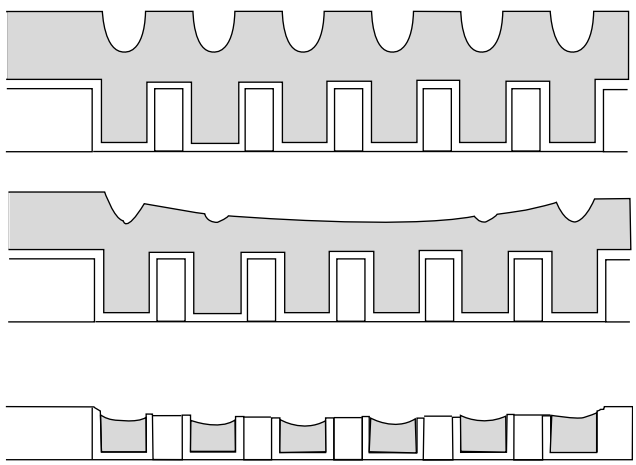


FIGURE 17.24 Schematic showing effect of dishing and erosion on a dense metal pattern.

the breakthrough phase of the polish to decrease the removal rate of both the field area and patterned regions. One can also see from Equation 17.27 that the erosion rate not only depends on the down force, platen speed, and pattern density, but also upon the consumables and dielectric material itself.

The heavy dependence of dishing on line width can be explained in terms of the pad deformation into the recessed areas. Since the metal polish rate is faster than the dielectric, the metal begins to dish and the pressure of the pad on the surface of the metal begins to decrease. In the case of infinite selectivity to dielectric, the pressure on the metal would soon begin to vanish resulting in a maximum amount of dishing, d_{max}. For a finite selectivity, dishing $d < d_{max}$ is observed which can be approximated [98].

$$d \sim d_{max}(1 - (K_{dielectric}/K_{metal})1/(1 - \rho)) \tag{17.28}$$

where K is the Preston constant for the different materials. This equation is for a specific line width and assumes an equilibrium condition that is independent of time. The pattern density relationship with dishing shows the dishing to decrease as the pattern density increases. This can be explained by the erosion mechanism: the removal rate of the dielectric increases with ρ, thus the maximum dishing is less. Although the dishing is less, the percentage of metal remaining is less (dishing + erosion) than isolated lines of the same dimensions. One must also note that in large patterned areas, the polish rate of the metal is faster due to the oxide erosion mechanism. As seen in Figure 17.24, the pattern density may also affect the metal removal rate over areas where pinch off of the metal has not occurred after deposition. This in turn clears the metal over large patterned areas faster than over field areas, thus leading to pronounced dishing and erosion. One would ideally like to have global planarization of the metal film prior to contact of the polish stop layer to reduce this enhanced dishing and erosion.

Although physical dimensions of dishing and erosion are constantly monitored in a metal CMP process, electrical impact is also of concern and needs to be managed [99]. A further concern beyond the first level metal is the effect of stacked metal layers, where the next dielectric layer and metal layer will fill or pool where the previous dished metal resides resulting in an unwanted bridging of metal.

17.3.3 Chemical Effects in Polishing

As shown above, models incorporating the Hertzian indenter mechanism can forecast the impact of pressure on removal rate. The shear tensor model predicts the impact of platen velocity on removal rate. Spatial averaging models can forecast step heights and thus the WIDNU. Use of pattern density test masks and the spatial averaging model generates a planarization length, which aids in process material and tool parameter design. These concepts were derived knowing only mechanical properties of the pad and the film on the wafer surface, viz. E_p and E_w. In this section we will examine how chemical effects can be used to modify the film's elastic modulus. This knowledge allows us to extend the model generated with homogeneous oxide films to homogeneous metal films. Ultimately, local damascene structures pattern loss must consider the corrosion potentials of mixed metal surfaces in contact with a good electrolyte, the slurry.

17.3.3.1 Chemical Effects in Oxide Dielectric CMP

In the literature, the chemical effect has been studied extensively by the glass polishing community. Cook [100] provides a detailed review of the chemical process in glass polishing. In general, the reactions between siloxane bonds (Si–O–Si) and water primarily determine the behavior of silicate surfaces during oxide polishing. The attack of the siloxane network will control the polish rate. This reaction can be described as a reverse depolymerization reaction, which can be expressed as:

$$(SiO_2)_x + 2H_2O \Leftrightarrow (SiO_2)_{x-1} + Si(OH)_4 \tag{17.29}$$

As suggested in Equation 17.29, water entry into the oxide surface weakens the glass network by breaking Si–O bonds. As a result, the hardness of the oxide surface is reduced by the hydration reaction. The evidence to support this theory is the polish rate of glass in oil and in dry medium (i.e., abrasive

only). Izumitani [1] found that the glass polish rate is substantially lower in oil and dry polishing media than in water. In oil and dry polishing, the polishing mechanism is the mechanical wear only (because the chemical mechanism due to water is absence), so that the polish rate is proportional to the hardness of the non-leached glass, and the rate is substantially lower than with a chemical plus wear mechanism. Note that the hydration reaction only occurs at the oxide surface. In practice, it is important to estimate the penetration depth of water in oxide surface (i.e., the distance that water diffuses into oxide). Cook calculated the penetration depth for water as a function of the diameter of abrasive particle and pressure [100]. At the glass/particle interface with polish pressures of 2 kg/cm^2, he estimated water penetration depths into the glass surface of 4 nm for a 0.5 μm-diameter particle. The water penetration depth increases with particle size and rises to 12 nm for a 5 μm-diameter particle at the same pressure. As a result, at least 10 nm of oxide needs to be removed in the post-CMP clean to eliminate the surface damage layer.

Nogami and Tomozawa [101] measured water diffusion in silica as a function of hydrostatic pressure and applied uniaxial stress. The diffusion coefficient increases exponentially with increasing tensile stress and decreases exponentially with increasing compressive stress and hydrostatic pressure. The solubility of water showed an opposite trend. As the abrasive particle moves across the surface, $Si(OH)_4$ solubility is high in front of the particle, while condensation dissolution occurs in back of the particle. Net material removal results only when the dissolution rate is greater than the condensation rate. There are five reaction steps important in determining the rate of mass transport during polishing. They are:

1. Water diffuses into the oxide surface
2. Water reacts with the surface, leading to dissolution under the influence of applied load
3. Some dissolution products adsorb onto the abrasive particle and are moved away from the surface
4. Some dissolution products redeposit back onto the surface
5. Surface dissolution occurs between particle impacts.

Thus, Cook views the chemical effect on oxide as kinetically generated. The stress of particle passage causes water to be "pumped into and out of" the oxide surface. This pumping action increases the frequency of exchange implied by Equation 17.29. This kinetic enhancement increases the solubility of the glass surface.

Another important factor in determining the polishing characteristics is the consumable set; that is, slurry constituent and polishing pad. Slurry chemistry (pH, anions, cations, and their concentrations) as well as slurry abrasive particles (size, solid content, type, and dispersion) can have strong effects on the polish rate, see Section 17.3 for details. In glass polishing, CeO_2 is often used as the abrasive particle. The slurry using CeO_2 as the abrasive particle has approximately three times higher polish rate than that using SiO_2. Jairath et al. [35] examined the use of CeO_2 as the abrasive particle in oxide CMP. Under the same polishing conditions, the CeO_2-based slurry has worse planarity results than the SiO_2-based slurry. It requires more oxide removal for the CeO_2-based slurry to reach planarity, even though the polish rate for CeO_2-based slurry is substantially higher than the SiO_2-based slurry. This is why the SiO_2-based slurry is widely used in microelectronics industry for dielectric CMP. Jairath et al. also investigated the effect of size and solid content of the abrasive particle on polish rate. It was found that the polish rate increases with both size and solid content of abrasive particles. However, such an observation is not consistent with the previous work in glass polishing. It was found that the polish rate is independent or decreases with abrasive particle size. Although a model based on the wear mechanism has been proposed in the literature, the effect of size and solid content of the abrasive particle has not yet been included in this model [72]. More work is required to explain the effect of size and solid content of the abrasive particle on the polishing characteristics.

17.3.3.2 Chemical Models in Tungsten CMP

In order to achieve a large process margin, i.e., stopping ability, when polishing damascene structures, it is important to have large polish rate selectivity of the inlaid metal to the dielectric. Tungsten is six

times harder than silicon oxides and aluminum, $E_w = 59$ Mpsi vs. $E_{SiO2} \sim E_{Al} = 10$ Mpsi [102]. Based on the Preston equation, Equation 17.4, for a non-selective slurry, i.e., pH ~ 7 and silica abrasive, the removal rate of silicon oxide should be six times faster than tungsten. This simple mechanical model is insufficient for metal CMP. The first CMP application on the front side of the SC wafer was to create tungsten plug contacts [4]. Metal CMP has even more chemical kinetics involvement than oxide CMP. An important concept in all metal CMP is the formation of a passivation or etching inhibition film in the down area of any pattern. Kaufman and co-workers [47] were the first to propose this passivation layer for tungsten CMP. The Kaufman model postulates a three-step mechanism to explain the planarization of tungsten under slurry conditions where the metal can be passivated:

1. A passivation layer forms rapidly on up and down areas of the metal due to the oxidation potential of the solution.
2. This freshly formed oxide film is softer than the metal itself and is mechanically removed from the up areas by the pad/abrasive, while the down areas are not contacted due to the partial flexibility of the pad.
3. The corrosive nature of the slurry re-grows the passive film on the up area, while the corrosive elements are blocked from the metal in the down area by the passivation film.

The up areas are continually corroded and polished until the pad flexibility allows it to contact the down areas. At that point, the planarization process again depends upon the hardness of the pad and the down force applied to the wafer. Thus, the corrosion properties of the slurry are chosen to convert the metal surface into a film which is soft enough to be abraded by the pad plus abrasive material, yet hard and impermeable enough to prevent outright etching of the metal. This is a simple model; the process is kinetically very complicated, yet the simple model is instructive because it utilizes the pad mechanics, that have already been developed for planarizing oxide on patterned wafers, to explain the planarization of a patterned metal surface. Creation of a planar surface, in the metal layer, before breakthrough to the underlying layers, i.e., "pre-planarization," is the second most important concept in metal damascene CMP processes. Because, if pre-planarization has not occurred at breakthrough, the corrosive nature of the slurry continues to attack the metal surface in preference to the dielectric, or the barrier layer in the case of copper or aluminum, which leads to dishing of the softer, metal areas. This dishing and erosion leads to large local and WIWNU.

Tungsten, being one of the refractory metals, tends to form strong oxygen bonds under oxidizing conditions. This oxide film is so dense on refractory metals that the metal is protected from further chemical attack and the metal is "passivated." Pourbaix documented the thermodynamic basis of passivation, for most elements of the periodic table, in his atlas [103]. The elements' solubility data is summarized in the form of "equilibrium diagrams," i.e., contour maps in a solution potential vs. pH field. Concentration contour lines from 1 to 1×10^{-6} mol/L are shown in $\log_{10}$ units. The region of high solubility, i.e., potential and pH are correct for high etch rates, are beyond the "0" contour line(s). The passivation region is found within or beyond the "−6" contour line(s). The tungsten equilibrium is shown in Figure 17.25. Tungsten is unique in that it can exist in solution only as an anion; the metal oxides are acidic. The passive region for the highest oxidation state, tungsten (+6), is below pH 4 and above −0.25 V vs. the normal hydrogen electrode (NHE). The lower oxidation states form stable passive layers at even more negative solution potentials and high pH's. Thus, on a thermodynamic basis, we understand why tungsten CMP slurries should be pH 5 or less. Again kinetics can modify the practical pH value higher or lower. In this case, favorable rates of formation of even more stable lower oxidation states has allowed slurry formulations up to pH 5 to be successful.

What about solution potential effects? Based on the favorable lower solubility of WO_2 and W_2O_5 in the potential range from −0.25 to −0.6 V vs. NHE, would it not suggest an oxidizer in this range would be a better choice than stronger oxidizing agents? A comparative study of just the oxidizer effect is not available in the open literature. A complication, due to favoring the higher oxidation state by using the iodate ion, IO_3^-, $E^0 = 1.2$ V vs. NHE, has been pointed out by Osseo-Asare et al. [104]. They have shown

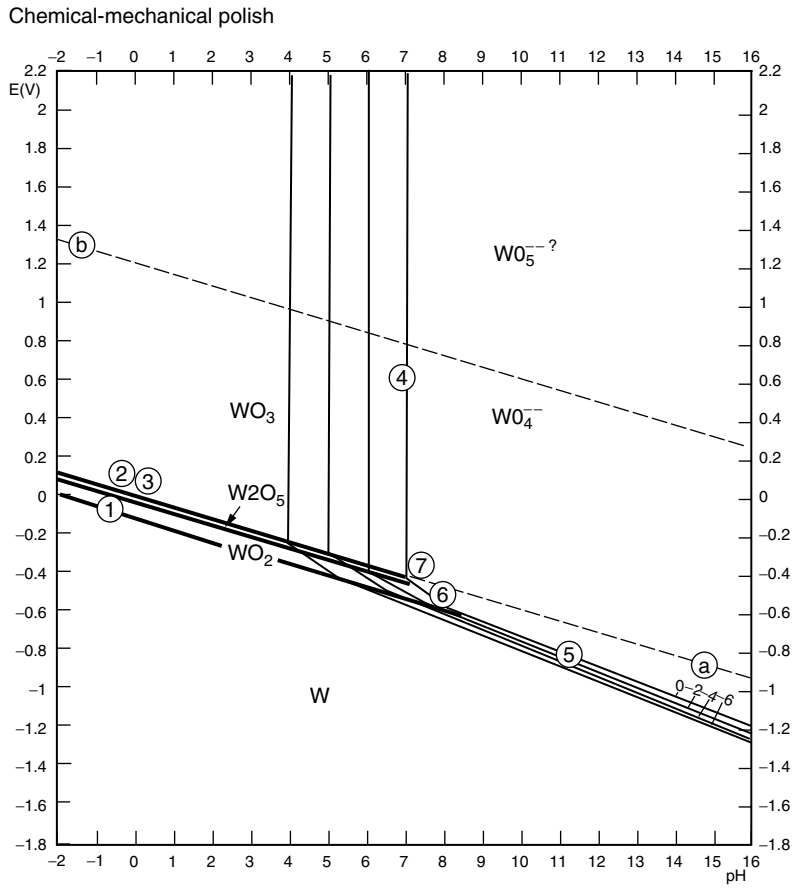


FIGURE 17.25 Tungsten solubility vs. solution potential-pH diagram. (From Pourbaix, M., *Atlas of Electrochemical Equilibria in Aqueous Solutions*, National Association of Corrosion Engineers, Houston, TX, 1974.)

increased anodic tungsten corrosion currents in the presence of increasing concentrations of tungstate ion, 1–10 mM, at pH 7. Also, at pH 4, the corrosion potential shifts more negatively in the presence of 10 mM iodate, when 10 mM tungstate ion is added. They suggest this enhanced dissolution rate of tungsten metal may be explained by the formation of polymeric species of tungstate ion:

$$W + 5WO_4^{-2} + H_2O = W_6O_{20}(OH)^{-5} + H + (+6e^-) \tag{17.30}$$

i.e., an autocatalytic polymerization reaction. Although more work is needed to confirm this hypothesis, it is suggestive of some practical solutions to problems that have been observed by tungsten CMP process engineers in the past. When using pads with high capacity to carry slurry, e.g., felts, the removal rate increases with number of wafers processed in each new lot, after a wet idle, up to some higher steady state value. This increase could be paralleling a build up of tungstate ion in the pad. Also, "keyhole" formation in the center of tungsten via plugs appears to have been solved by using closed celled, molded polyurethane pads, instead of felt pads, and hydrogen peroxide, instead of iodate or ferric ions as oxidizer in the slurry. Could this also be due to a reduction of the quantity of tungstate ion produced by hydrogen peroxide and carried on the less porous pads?

17.3.3.3 Mechanisms in Copper CMP

Copper is softer than the barrier metals, and other liner materials, but harder than aluminum and silicon oxide, $E_{Cu} = 19.2$ Mpsi, $E_W = 59.4$ Mpsi [102]. Under purely mechanical abrasion conditions, the Preston equation, Equation 17.4, predicts copper should polish three times faster than tungsten. One must never discount the possibility that some polishing can occur by purely mechanical abrasion of the metal itself. Steigerwald et al. [8] have determined copper polish rates with non-patterned wafers on Suba IV pads. Slurry composed of 2.5% alumina and 2% ammonium hydroxide, at flow rates of 250 ml/min was used. At pad velocities of 130 cm/s, they found the copper polish rate, 300–800 nm/min, linearly dependent on pressure in the range 12–21 kPa. At carrier pressures of 15 kPa, they found the copper polish rate, 200–950 nm/min, linearly dependent on pad velocity in the range 4–22 cm/s. On harder Suba 500 and Suba 550 pads, at 15 kPa carrier pressure, the copper polish rates became sub-linear at approximately 600 nm/min and pad velocities of 130 and 12 cm/s for Suba 500 and Suba 550, respectively. The Preston constant was calculated from the slope of linear portion of the curve. The RR-velocity curves gave a value of 5×10^{-13} Pa^{-1} and the RR-pressure curve gave a value of 4×10^{-13} Pa^{-1}. The theoretical Preston constant, $K_P = 1/2E_{Cu}$, is 3.8×10^{-12} Pa^{-1}. Finding the actual proportionality constant to be 10%–12% of the theoretical value is typical. The important point is that the Preston relationship is valid for a metal like copper, under constant chemical conditions, low carrier pressures, and low pad velocities.

The passivation of copper is more complex than tungsten, as shown by Figure 17.26 [103]. In the absence of complexing agents, e.g., ammonia, organic amines, cyanide, halogens, thiocyanide, etc., the cupric oxide form is stable between pH 7 and pH 12.7. Cuprous oxide, Cu_2O, is stable from pH 5.5 to pH 14, at potentials of 0.1 V vs. NHE or less. Using x-ray photoelectron spectroscopy it has been shown, under static etching conditions with slurries from pH 5 to 11, and polishing with these slurries, the dominant copper oxidation state is the cuprous, i.e., Cu^{+1} [8]. Steigerwald et al. [8] did a thorough study of the electrochemical control of copper dissolution in ammonium hydroxide base slurries. He points out that the dynamic nature of the CMP process may favor the formation of cuprous oxide as the intermediate passivation film. Furthermore, he postulates the cuprous ion, as the amine complex, is the first soluble intermediate to enter the slurry. Formation of this species then, is the rate-determining step in ammonia based slurries. The effect of various ammonia slurries on the mixed copper electrode potential, its change during polishing of a non-patterned copper wafer and the average polish rate are shown in Table 17.4. The mixed electrode potential and pH of the ammonium nitrate slurry would predict that the stable copper species be Cu^{+2}. For ammonium chloride slurry Cu metal and for ammonium hydroxide Cu_2O would be the stable species [8]. However, Johnson and Leja [105] have recalculated the stable species as a function of electrode potential and pH, in the presence of ammonia, and found that the stable species for these slurries would be Cu for ammonium chloride and Cu $(NH_3)_4^+$ for both ammonium nitrate and ammonium hydroxide. Examining the table, we see the electrode potential of the ammonium nitrate and ammonium hydroxide based slurries change positively before and after polishing, whereas the ammonium chloride slurry's potential does not change. This increased electrode potential is indicative of an oxidized form of copper in the slurry after polishing the copper surface. Notice that the ammonium chloride slurry polish rate is 1/3 and 1/6 of the nitrate and hydroxide slurries' rates. Thus the polish rates are faster when a stable soluble species can be formed.

As long as the ammonia-based slurry is discarded after a single pass, i.e., not re-circulated back to the polish pad, ammonia initiated corrosion problems will not occur. However, the cuprous ion is not only stabilized and made more soluble by complexing agents, particularly ammonia and amines, but it is also quickly air oxidized to the complexed cupric ion at pH >7:

$$2Cu(NH_3)_4^+ + H_2O + 1/2O_2 = 2Cu(NH_3)_4^{+2} + 2OH^-, \quad E^0 = 0.94\ V \tag{17.31}$$

$$Cu + Cu(NH_3)_4^{+2} = 2Cu(NH_3)_2^+, \quad E^0 = 0.39\ V \tag{17.32}$$

As Equation 17.32 shows, the cupric tetraammonia ion can quickly dissolve the copper metal by these reactions. Although these reactions would only happen in the pH range 7–11, similar reactions can be written on the acid side of pH 7 with bromide ion replacing ammonia as the ligand. Thus, in order to

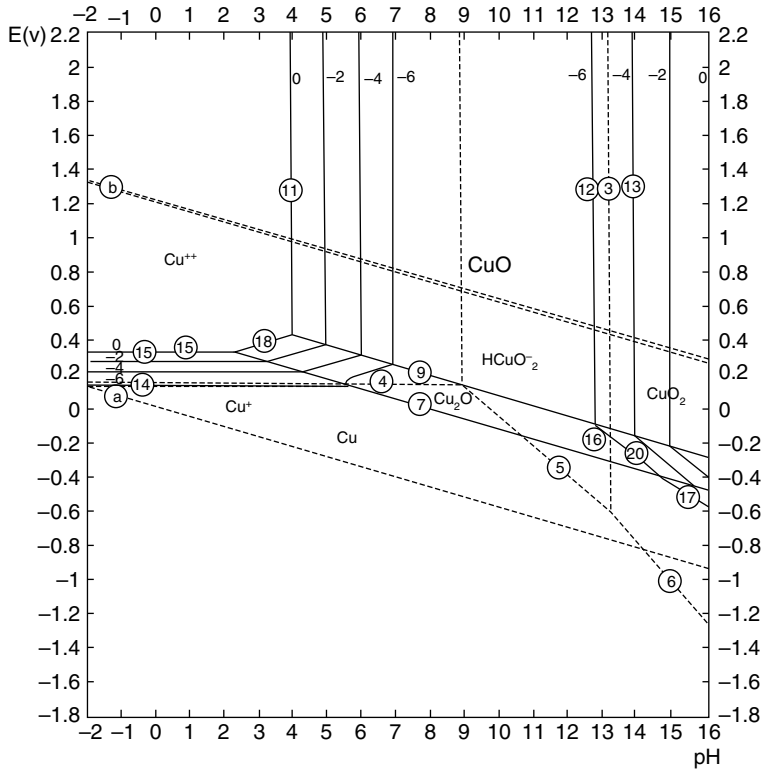


FIGURE 17.26 Copper solubility vs. solution potential-pH diagram. (From Pourbaix, M., *Atlas of Electrochemical Equilibria in Aqueous Solutions*, National Association of Corrosion Engineers, Houston, TX, 1974.)

avoid unwanted corrosion reactions turning the polishing slurry into one with a high etch rate; build up of cuprous and cupric ion in the slurry is to be avoided.

In an attempt to increase the copper removal rate, Steigerwald et al. [8] have successfully polished copper with an alumina–ammonia-based slurry using potassium ferricyanide as the oxidant. They were able to demonstrate planarization with patterned wafers. The polish rate, on non-pattern wafers, was found to be 2600 nm/s and the etch rate was only 2 nm/s. Thus, the cuprous ferricyanide acts as an excellent passivation layer in the down area, and the Kaufman model for metal planarization is also demonstrated for copper.

The use of pH 11 ammonia-based slurries is only of academic interest because the silicon oxide dielectric would polish very rapidly with alumina as well as silica-based slurries at this pH. Thus the

TABLE 17.4 Effect of pH in Ammonia Slurries on Copper Polish Rate

Chemical	Conc., All= 0.167 M NH_3 (%)	pH	E, mv vs. NHE	ΔE, mv $\pm$ range	Polish Rate,[a] nm/min
NN_4NO_3	1.4 wt.	4.7	220	70 $\pm$ 50	175 $\pm$ 50
NN_4Cl	0.9 wt.	4.8	60	0.0 $\pm$ 25	50 $\pm$ 50
NN_4OH	1.0 vol.	11.0	−80	125 $\pm$ 75	310 $\pm$ 90

[a] Strasbaugh 6CU polisher, Suba IV pads, 2.5% alumina, $P = 3$ kPa, $V = 110$ cm/s.

Source: From Steigerwald, J. M., Zirpoli, R., Murarka, S. P., Price, D., and Gutmann, R. J., *J. Electrochem. Soc.*, 141, 2842, 1994.

TABLE 17.5 Oxidants for Copper Chemical–Mechanical Planarization or Chemical–Mechanical Polishing (CMP) Slurries

Redox Reactions	E^0, Volts vs. NHE
$H_2O_2 + 2H^+ + 2e^- \leftrightarrow 2H_2O$	1.776
$MnO_2 + 4H^+ + 2e^- \leftrightarrow Mn^{2+} + 2H_2O$	1.224
$IO_3^- + 6H^+ + 6e^- \leftrightarrow I^- + 3H_2O$	1.085
$HO_2^- + H_2O + 2e^- \leftrightarrow 3OH^-$	0.878
$Fe(CN)_6^{-3} + e^- \leftrightarrow Fe(CN)_6^{-4}$	0.358
$IO_3^- + 6H_2O + 6e^- \leftrightarrow I^- + 6OH^-$	0.260

damascene process would not achieve high copper to dielectric polishing selectivity, which was stated in the previous section as the first principle of damascene CMP. In order to increase metal to silicon oxide selectivity, pH's of 7 or less are usually used. Because ferricyanide reduction involves electron exchange, without water–proton exchange, the redox couple is independent of pH, see Table 17.5. Thus this oxidant could be used at pH 7 or less. However, the ferricyanide radical is less stable in low pH solutions and it liberates hydrogen cyanide, which is poisonous. Bromide and chloride can be used as cuprous ion passive film forming agent in the ideal pH range of 4–7. In addition, the continued oxidation of the halogen complexed cuprous ion to the halogen complexed cupric ion provides a mechanism for dissolution of the polished residue in the used slurry. They should be the ideal passive film-forming, ligand-forming anions, except for the corrosion reaction already discussed, see Equation 17.32. Comparing Table 17.5 oxidants with Figure 17.26 demonstrates that all are thermodynamically capable of oxidizing copper to the cupric ion. Manganese dioxide is particularly interesting because it is a solid and, properly manufactured, it may also function as the slurry abrasive as well as the oxidant. Care must be exercised in choosing the solublizing ligand for this oxidant. It is strong enough to react explosively with organic alcohols and the halide ions are strong reducing agents!

Patterned copper wafers can also be planarized using the corrosion inhibition mechanism when BTA is present in the slurry [51–53]. Several papers have suggested formulations to make practical slurries based purely on this corrosion inhibition agents, surface adhesion to copper metal and its ligand formation with the copper ions [106,107]. No practical slurries, based on BTA, an abrasive and an oxidizing etchant alone are commercially available. Benzotriazole is mainly used as a slurry additive, to inhibit the tarnishing of the freshly polished copper surface. Corrosion/tarnish inhibition is critical while copper is still in contact with the spent slurry on the polish pad, before adequate dilution and flushing with DI water.

17.4 Applications and Issues

Several of the more common applications for CMP in integrated circuit manufacturing are discussed in Section 17.4.1. The focus is on practical issues for implementation of these CMP processes.

17.4.1 Dielectric CMP Applications

17.4.1.1 Shallow Trench Isolation (STI)

Shallow trench isolation has replaced local oxidation of silicon (LOCOS) as the preferred technique for isolation of active areas in silicon integrated circuits. Shallow trench isolation has the advantages of a more planar structure and a greater circuit density since it eliminates the problems associated with the encroachment of the isolation oxide "bird's beak" into the active regions [108]. Chemical–mechanical planarization or chemical–mechanical polishing has been used to reduce topography introduced during

LOCOS isolation, but the oxide encroachment remained. In the STI process sequence, a trench is etched into a silicon nitride layer and into the silicon substrate. The trench is filled with a thin liner oxide and then a deposited oxide such that the fill oxide level is slightly above the original silicon level. The oxide over burden is removed by an oxide CMP process. In some cases, a combination of a plasma etch-back followed by oxide CMP process is used. The silicon nitride layer acts as a polish stop to protect the underlying active regions from CMP damage and is later removed using a wet chemical etch, typically in hot phosphoric acid.

Chemical–mechanical planarization or chemical–mechanical polishing pattern effects (see Section 17.3.2) lead to a variation in the oxide removal rate as a function of the local pattern density. Sufficient polish time must be given to clear oxide above nitride in all regions of the die or there will be residual nitride remaining after the nitride strip process. Residual nitride will result in transistor failure. However, over-polish of the wafer can lead to problems with thinning of oxide in wide features due to dishing [109]. More importantly, the low selectivity of typical oxide CMP processes using standard fumed or colloidal silica slurries leads to thinning of oxide and erosion of silicon nitride in less dense regions and can lead to silicon damage in what will be an active gate region. In general, the STI CMP process window time is very tight and therefore active endpoint control is desirable. The process requires uniform oxide removal across the wafer, as uniform a pattern density as possible within the die to reduce over-polish requirements, and preferably a stiff pad and optimized CMP process conditions to minimize dishing topography.

Complicated and costly process integration sequences that include patterning and etch-back of excess oxide prior to CMP have been used to improve the process window for STI CMP. Patterned resist etch-back and direct reactive ion etch RIE of the oxide over large or high density features have been reported [110–114]. The removal of oxide in regions that would polish more slowly during CMP results in a wider process time window. The overall STI module integration requires optimization of deposited oxide thickness, oxide type, pattern sizing, and etch-back duration. The effective pattern density variation within the die can be improved by insertion of dummy active features in less dense regions. Another approach that has been reported to reduce dishing is the deposition of a thin polish stop layer such as nitride over the oxide after etch-back [115]. The polish-stop layer over high features is readily removed due to the local high polish pressure, but slows dishing in wide features.

Direct polish of the oxide without a patterning step is desirable to reduce manufacturing costs, but difficult to achieve in practice with standard silica slurry formulations. One approach using a silicon nitride over layer and a one-step CMP process at high speed and low pressure was reported to have achieved STI planarization without excessive nitride thinning or wide field dishing [116]. Use of a nitride cap layer has been shown to also reduce the number of micro-scratches in the field oxide [117]. Other materials with a higher oxide to over layer selectivity have been used. One group reported using boron nitride, with oxide to BN selectivity above 30:1 [118]. Development of new commercial slurries with higher oxide to silicon nitride selectivity has also been achieved [119,120]. Ceria suspensions are being evaluated because of the higher selectivity attainable, but higher levels of scratch defects are often observed. Selectivity measured on blanket wafers is not always representative of what is achieved on patterned wafers. Successful implementation of a direct CMP process for STI will require high selectivity on patterned wafers, a slurry and process that is self-limiting in oxide removal rate once planarization occurs, and design rules that lead to uniform pattern densities.

17.4.1.2 ILD Planarization

Dielectric CMP is needed to remove the topography introduced in etched polysilicon gate and etched metal process flows. Chemical–mechanical planarization or chemical–mechanical polishing has replaced re-flow of doped glasses, spin-on-glass etch-back and resist etch-back because these other techniques only achieve local planarization. Unlike CMP for STI applications, there is typically not a polish stop layer, and repeatable control of polish rate and uniformity is required. Process variations such as pad age, slurry batch changes, machine calibration drift, pad conditioner age, and chamber-to-chamber dielectric film properties variation can all result in significant variations in removal rate and process uniformity.

Repeatable process results require that the pad be conditioned to maintain a consistent pad surface. Pad conditioning can be done in between wafers or during wafer polish. The CMP process engineer has relied heavily on external film thickness metrology for frequent tool and product qualification tests. However, the post-CMP clean and ex situ metrology operations are time consuming and reduce useful CMP tool availability. In-line and in situ metrology tools are now available for more rapid feedback or real-time feedback for process control of the CMP tool [95].

Pattern dependent CMP rates of the dielectric lead to thickness variations within the die that exceed across wafer variations, see Section 17.3.2. This thickness variation can lead to problems with via etch or contact etch and yield loss. Dummy poly silicon or metal structures can be inserted to tighten the pattern density distribution and can help to reduce the final within-die thickness variation [121]. However, the benefits from improved ILD thickness must be balanced by the potential loss in circuit performance due to the increase in parasitic capacitance of active circuit to dummy features. A decision on whether to allow the dummy features to float electrically or connect them to ground (or Vcc) must be made based on circuit performance and reliability concerns. Another pattern density issue to consider is the presence of large metal areas near the edge of the wafer in non-patterned areas. This results in areas of high pattern density, which lower the CMP removal rate in the adjacent edge die. Patterning of partial die at the wafer edge will improve dielectric thickness uniformity, but at the expense of photolithography tool capacity [122].

Chemical–mechanical planarization or chemical–mechanical polishing process conditions and consumables also affect within-die oxide thickness variation. Pad deformation is greater at high pressure/low speed conditions than it is at low pressure/high speed conditions, which leads to a shorter planarization length. Pad properties also have a large effect on planarization. The industry standard pad for dielectric CMP has been a stacked IC-1000/SUBA IV from Rohm and Haas, which consists of a hard cast polyurethane pad on top of a soft, compressible felt pad. The hard top pad is needed for planarization and the soft pad underneath has improved across wafer thickness uniformity compared to polishing with not sub-pad. However, the underlying soft pad allows greater pad stack compressibility and degrades planarization capability. Total pad stack compressibility can be changed by sub-pad material selection or varying the relative thickness of the two layers. The effect of process conditions on planarization length is small compared to the improvement found when changing from the stacked pad to an IC-1000 pad with no sub-pad [123].

17.4.1.3 Polysilicon CMP Applications

Chemical–mechanical planarization or chemical–mechanical polishing is also used for several other non-metal applications in IC manufacturing. Polysilicon deep-trench capacitors are formed using CMP to etch-back the polysilicon fill material from the wafer [109]. With CMP, issues with center seam propagation often seen with plasma etch processes are eliminated. The polysilicon slurries are highly selective to the oxide or nitride hard-mask materials and feature sizes are small, therefore the process is self-stopping and has a relatively wider process time window compared to ILD or STI CMP applications. Dishing is not of a great concern because of the typically small dimensions, but dielectric erosion can be a concern in regions of high pattern density. Other polysilicon CMP applications include the formation of polysilicon plugs for isolation, trenches of polysilicon local interconnect, and the use of CMP to smooth blanket polysilicon films prior to subsequent processing.

17.4.2 Metal CMP Applications

One key advantage of metal CMP over oxide CMP is that the commercially available in situ end pointing methods are more sensitive for metal CMP. Due to the heterogeneity of the metal CMP process one can endpoint either by monitoring temperature, reflectivity, acoustically, chemically, or torque of the platens or carriers due to the different frictional coefficients of the materials as discussed in previous sections. One would like to have a clear endpoint signal followed by a short amount of over-polish due to the reasons stated in Section 17.3.2. An example of a clear endpoint signal can be seen in Figure 17.10. This signal shows

a clear endpoint due to the reflectivity difference of the metal vs. the reflectivity of the polished patterned wafer. To obtain clear endpoint signals one must develop a process where the wafer clears uniformly resulting in an unambiguous signal, in spite of sampling a small portion of the wafer.

Although there are many different issues between Cu, Al, and W CMP such as incoming material (PVD, CVD, electroplated), slurries (PH, abrasive, etc.), and pads (compressibility, grooves, pore size, etc.), physical vapor deposition the goals of metal CMP are the same. This is to remove metal and barrier off the surface of the dielectric while minimizing the erosion of the dielectric and to minimize the dishing of the metal in the patterned regions while leaving both surfaces free of defects and surface roughness.

17.4.2.1 Tungsten CMP

Tungsten is primarily used at via levels and most notably at the contact level due to its chemical stability and electromigration properties [124]. Because most tungsten films are deposited by CVD techniques, pinch off can occur during via filling. This leaves a seam that must be tightly controlled in order to prevent key holing during CMP polish or tungsten etch back. One would like slurry with a low static etch rate to prevent chemical attack preferentially down the seams of the tungsten. In Figure 17.27, one can see the tungsten seam and the key holing of the tungsten that can occur. One must be sure that the tungsten seam is very minute after deposition so the slurry cannot get into the seam.

17.4.2.2 Copper CMP

Copper CMP is much less mature than tungsten CMP but many of the same issues exist. Many additional issues are added to copper CMP due to the material itself. Although copper is more noble than tungsten, the metal oxide layer is much more permeable to chemical attack and further oxidation [125]. This in turn limits the materials in contact with copper, which may induce corrosion of the copper metal. Some fixes to corrosion involve the addition of an inhibitor such as BTA (1 H–benzotriazole) to bind the surface and limit the amount of electrochemical attack at the surface during rinsing of the wafer or added to the CMP slurry [126]. Most processes involving copper CMP have utilized a dual damascene approach as mentioned in previous sections. This simplifies the flow for creating interconnect but complicates the control of the Cu CMP process. One would like to have inline metrology to monitor the process such as

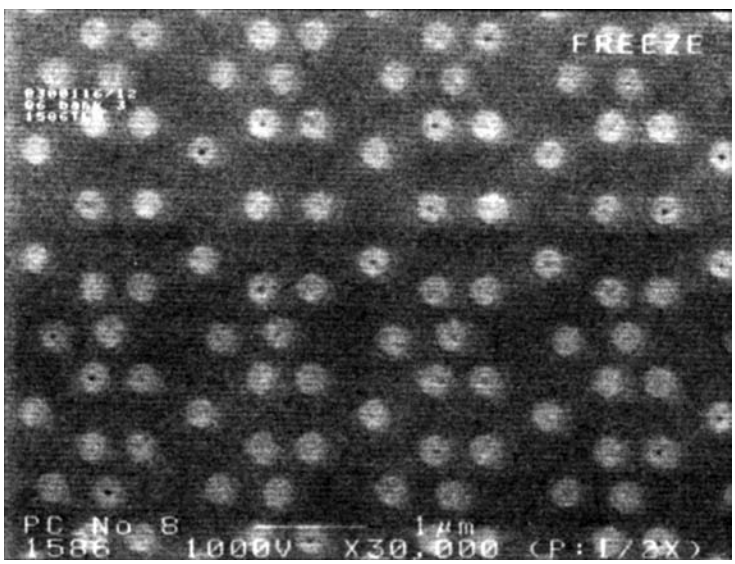


FIGURE 17.27 Top down SEM view of an array of tungsten plugs, after CMP, revealing the see due to tungsten chemical vapor desposition.

measurements of the oxide thickness (erosion) to monitor the process. Since dual damascene involves filling the via and trench at the same time, one cannot create a buried block of metal at the via level to use as a substrate for post-CMP dielectric thickness measurements that would monitor erosion. Dishing measurements are possible but are not production worthy due to the time requirements and subjective leveling and cursor placement. One is limited to the inline probe to monitor processes, which usually is further down the process flow, limiting the ability of the Cu CMP engineer to make quick fixes to problems. Other issues for Cu CMP process and process integration are self-annealing [127] and grain orientation [128] causing CMP removal rate variation, pattern density, electroplated copper chemistry, annealing, slurry selectivity, and cleaning chemistry limitations.

17.4.2.3 Aluminum CMP

Due to the mature process of depositing aluminum, some fabs have interest in incorporating a damascene approach with this material. This is mainly to replace the higher resistive metal tungsten with the more conductive aluminum [129]. Also the dual damascene approach leads to much lower resistivity lines due to elimination of contact resistance between dissimilar metals [130]. Aluminum CMP is also in the infancy stage as with copper CMP, and many of the same issues one faces with copper, one will also face with aluminum. Aluminum, being softer than copper, tends to scratch and dish more than copper [131]. Aluminum also forms the impervious Al_2O_3 layer very readily upon exposure to air and or oxidizing chemicals. The aluminum oxide layer is much harder than the aluminum and removal during CMP by more mechanical means than chemical leads to surface scratches and roughness on the aluminum surface. Although the aluminum oxide layer is present on the surface, the film passivates the surface of the aluminum and further corrosion or chemical attack is much less likely than in copper. One objective in aluminum CMP is to have the slurry oxidize the surface of the metal at the same rate that one is mechanically removing this layer. This is to increase the removal rate of the aluminum and have the dissolution products be a softer hydrated aluminum oxide leaving the aluminum surface oxide free and smooth. Although aluminum CMP is not mature, more experience has been gained in the production of IC using aluminum as the interconnect material of choice.

17.5 Post-CMP Clean

17.5.1 Post-CMP Clean for Dielectrics

Since CMP is a wet process, CMP cleaning technologies are mostly aqueous. The post-CMP wafer should be kept wet and no slurry should be allowed to dry and permit strong adhesion of particles onto the wafer surface. In practice, post oxide CMP wafers are typically cleaned by the SC1 solution, i.e., the mixture of H_2O, H_2O_2, and NH_4OH. The best way to understand the particle cleaning is to look at the electrostatic interaction between particles and substrate. Zeta potential, the electrostatic potential a few angstroms away from the surface, is typically used to define the charge between particle and substrate in an aqueous environment. Figure 17.28 shows the ZP of silica, alumina, silicon nitride, and polyvinyl alcohol (PVA) [132]. Zeta potential is a strong function of the pH of a solution. If the ZP on the particles is large, then the solution pH must be adjusted such that the particle, wafer surface, and brush material all have like charges. As seen in Figure 17.28, the ZP for silica, alumina, and PVA become strongly negative at high pH; the addition of NH_4OH will increase the electric repulsive force between these materials. This is why SC1 type of solution is effective to remove the particles for oxide and tungsten surfaces. SC1 chemistry generally has high particle removal efficiency but low surface metal removal efficiency, while acidic baths, such as SC2, i.e., the mixture of H_2O, H_2O_2 and HCl, or dilute HF, remove surface metals but often re-deposit particles. As a result, SC1 is used first and acidic solutions (such as HF or SC2) are often used after SC1 in the post-CMP cleaning sequence.

Another way to increase particle removal efficiency is to add mechanical action to remove particles from the oxide surface. It has been widely reported in the literature that either brush scrubbing or

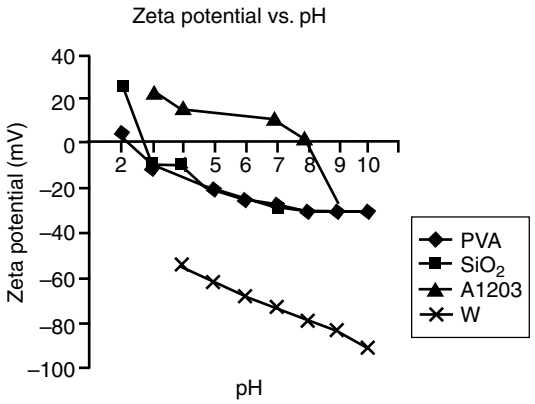


FIGURE 17.28 Zeta potential as a function of pH for polyvinyl alcohol, silica, alumina and tungsten. (From Malik, I. J., Zhang, J., Jensen, A. J., Farber, J. J., Krusell, W. C., Raghavan, S., Rajhunath, C., *Material Research Society Symposium*, 386, 1995.)

megasonic energy can provide the mechanical action for surface particle removal [133–135]. In addition, it has been proposed that post-CMP surfaces can be cleaned by super-cooled ice particles [136]. Both scrubber and megasonic cleaning are, however, the most popular techniques for post-CMP clean. Thus, we mainly focus on these two aqueous approaches in this section.

For the scrubber clean, PVA is the most widely used material. Unlike rigid brushes such as nylon, commonly used in the 1970s, PVA does not cause scratches when brought in contact with the wafer surface [137]. PVA material has an open structure consisting of interconnecting cells that allow the brushes to be consistently flushed with DI water during wafer cleaning [137]. These softer brushes clean more efficiently when they are compressed against the wafer, and higher brush speed can improve the particle removal efficiency without introducing scratches. In addition to mechanical scrubbing, commercial scrubbers typically provide SC1 solution to assist particle removal.

The major difference between megasonic and scrubber clean is that megasonic clean does not require a direct mechanical contact with wafer surfaces. Megasonic cleans were first presented by scientists at RCA [138]. The technique allowed them to use their SC1 effectively at lower temperature (30°C instead of 70°C). In megasonic clean, ultrasonic energy with frequencies near 1 MHz is added to the clean bath. The key for megasonic cleaning is the presence of a small boundary layer that essentially shields the surface. In the absence of megasonic agitation this boundary layer is hydrodynamically defined and relatively thick. Particles with sizes smaller than the thickness of the boundary layer are exposed to a much lower fluid velocity as compared to that outside the boundary layer. These lower velocities only provide a low removal force that may not be sufficient to remove particles. When the ultrasonic energy is introduced, the boundary layer becomes thinner leading to high particle removal efficiency. Fraser et al. reported that the megasonic clean is efficient enough to remove particles without using mechanical scrubbing [135].

Aside from the scrubber and megasonic cleans, the polishing process also plays an important role in achieving low particle levels. It has been reported that the secondary buff polish, i.e., polish on a soft pad and high pressure DI water spindle rinse at high spindle rotation after polish, are critical for particle reduction [139]. As a result, a second platen for buff polish is often used on commercial polishers. Ali et al. [140] reported that the pH in buff polish also affects the final defect counts. For DI water buff, the defect level decreases with an increase in pressure. However, when a basic solution is used on a secondary platen, pressure can have a deleterious effect on defect level, perhaps through pressure-induced coagulation.

In addition to the particle, metallic contamination is another issue for post-CMP clean. Since many commercial slurries are KOH-based, an effective clean is needed to remove surface metals, especially alkaline metals. It has been reported that removal of several 100 Å of top oxide surface is required in

order to remove alkaline metals from post-CMP surface [136]. As a result, an "HF deglaze" is usually employed to reduce metallic contamination and some embedded CMP defects. This is also an indirect evidence of the existence of a surface damage layer due to CMP [141–144].

17.5.2 Post-CMP Clean for Metals

The goal in post metal CMP is the same as for the post oxide CMP, namely to reduce defects. For metal CMP there are additional requirements for the post-CMP clean: it should not leave the surface of the metal with an oxide layer, it should leave the metal surface smooth, and it should not attack the barrier thereby leaving a recess around the metal. Although in metal CMP the slurries used vary widely in pH and abrasive, the cleaning chemistry is not much different. As discussed previously, the wafers must be kept wet prior to cleaning and then the wafers are subjected to chemistry whose pH is adjusted to be above the zeta potential of the abrasive particles and the surface of the wafer. This is usually under caustic conditions because the particles must be repelled from the surface of the wafer and one another to be freed from the surface [145,106]. One of the biggest challenges in post-CMP clean of metals is corrosion. If one is using ammonia for post-CMP clean of copper, one must be careful since it is known that ammonia will complex with copper and cause surface roughness and corrosion products [146].

Cleaning with high pH chemistry leaves the surface relatively free of particles but the metal contamination level may be higher than desired. Acid chemistries are employed in post-CMP clean of metal surfaces to reduce the metal contamination [147]. One must clean the front and back surfaces of the wafer to be metal free so as not to cause cross contamination throughout the fab and to improve device performance by removing the mobile ions and trace metals which may diffuse into the silicon. Tungsten and aluminum form a good passivation layer that is hard and etches slowly in mineral acids, while copper has a soft porous oxide layer through which chemicals can diffuse and cause further oxidation/corrosion. Protection of the copper surface while cleaning the dielectric surface free of metals may be accomplished by using inhibitors [126,148,149,150]. Other approaches may use short time clean in acidic environments, or by utilizing the oxidation and etching chemistries in one solution such as done with SC1 [151], or separately such as NH_4OH followed by HF.

Hydrofluoric acid is used in many fabs for two reasons: (a) to remove the surface of the dielectric which is embedded with metal ions, and (b) to help dislodge any particles that may have embedded into the dielectric. Although HF may seem advantageous, one must be careful with concentration and time used for metals decontamination [152]. Other than HF, one can use many forms of carboxylic acids or EDTA which chelate with the metals and reduce the presence of metals on the surface of the wafer [153]. While chemistries play a major role in the post-CMP clean, the polishing parameters such as pad hardness, buffing, and over-polish time also play a significant role in post-CMP clean. Although one can clean wafers fairly well within the polisher using soft pads and high pH chemistry, the CMP engineer is reluctant to sacrifice throughput by adding more time to the buff step. However, the engineer knows that most slurry particles should be left behind in the polisher by a rinse step or buff prior to the wafer entering a post-CMP cleaning tool.

References

1. Izumitani, Z. In *Treatise on Materials Science and Technology*, edited by M. Tomozawa, and R. Doremus, New York: Academic Press, 1979.
2. Sivaram, S., K. Monnig, R. Tolles, A. Maury, and R. Leggett. In *Overview of Planarization by Mechanical Polishing of Interlevel Dielectrics*, *Symposium on ULSI Science and Technology*. 3rd ed., 606–16. Pennington, NJ: The Electrochemical Society, 1991.
3. Head, N. L., R. D. Hempstead, and T. N. Kennedy. *Corrosion Resistant Thin Film Head Assembly and Method for Making*. U.S. Patent 4,130,847, December 19, 1978.
4. Fury, M. "The Early Days of CMP." *Solid State Technol.* 40-5 (1997): 81–6.

5. Brors, D. L., K. A. Monnig, J. A. Fair, W. Coney, and K. C. Saraswat. "CVD Tungsten—A Solution for the Poor Step Coverage and High Contact Resistance of Aluminum." *Solid State Technol.* 27 (1982): 313.
6. Chow, M.-F., W. L. Guthrie, and F. B. Kaufman. *Method of Forming Fine Conductive Lines, Patterns and Connectors*. U.S. Patent 4,702,792, October 27, 1987.
7. International Technology Roadmap for Semiconductors, 2005 edition, http://public.itrs.net, 2005.
8. Steigerwald, J. M., S. P. Murarka, and R. J. Gutmann. Chemical Mechanical Planarization of Microelectronic Materials. New York: Wiley, 1997.
9. Li, S. H., and R. O. Miller, eds. *Chemical Mechanical Polishing Silicon Processing*, *Semiconductors and Semimetals*, Vol. 63. San Diego, CA: Academic Press, 2000.
10. Oliver, M. R., ed. *Chemical–Mechanical Planarization of Semiconductor Materials*. New York: Springer, 2004.
11. Economikos, L., X. Wang, X. Sakamoto, P. Ong, M. Naujok, R. Knarr, L. Chen, et al. In *Integrated Electro-Chemical Mechanical Planarization (Ecmp) for Future Generation Device Technology*, *International Interconnect Technology Conference*, 233–5. 2004.
12. Wada, Y., I. Noji, I. Kobata, T. Kohama, A. Fukunaga, and M., Tsujimura. In *The Enabling Solution of Cu/Low-k Planarization Technology*, *International Interconnect Technology Conference*, 126–8. 2005.
13. Suni, I. I., and B. Du. "Cu Planarization for ULSI Processing by Electrochemical Methods: A Review." *IEEE Trans. Semicond. Manufact.* 18, no. 3 (2005): 341–9.
14. Preston, F. W. "The Theory and Design of Plate Glass Polishing Machines." *J. Soc. Glass Technol.* 11 (1927): 214–54.
15. Luo, J., and D. A. Dornfeld. "Material Removal Mechanism in Chemical Mechanical Polishing: Theory and Modeling." *IEEE Trans. Semicond. Manufact.* 14, no. 2 (2001): 112.
16. Jairath, R., A. Pant, T. Mallon, B. Withers, and W. Krusell. "Linear Planarization for CMP." *Solid State Technol.* 39, no. 10 (1996): 107.
17. Gotkis, Y., D. Schey, S. Alamgir, J. Yang, K. Holland. In *Cu CMP with Orbital Technology: Summary of the Experience*, *IEEE Advanced Semiconductor Manufacturing Conference*, 364. 1998.
18. Perry, K. In *Chemical Mechanical Polishing: the Impact of a New Technology on an Industry*, *IEEE Symposium VLSI Technology*, 2–5. 1998.
19. Chikaki. S., *Wafer Polishing Apparatus*. U.S. Patent 5,542,874, August 1996.
20. Ikenouchi, K., T. Murakami, and Y. Miyoshi. In *Conference Proceedings of CMP–MIC*. Santa Clara, CA, 271, 1999.
21. Film Backed Carrier Drawing is Provided Courtesy of Applied Materials 3111 Coronado Dr., Santa Clara, CA 95054.
22. Bladder Carrier Drawing is Provided Courtesy of Applied Materials 3111 Coronado Dr., Santa Clara, CA 95054.
23. Perlov, I., E. Gantvarg, and S.-H. Ko. *Carrier Head with a Flexible Membrane for a Chemical Mechanical Polishing System*. U.S. Patent 5,964,653, October 1999.
24. TBW Industries, Inc. Forest Grove Rd., Furlong, PA 18925.
25. ABT, Abrasive Technology, 8400 Green Meadows, Westernville, OH 43081.
26. Breivogel, J. R., L. R. Blanchard, and M. J. Prince. *Polishing Pad Conditioning Apparatus for Planarization Process*. U.S. Patent 5,216,843, June 8, 1993.
27. IPEC 472 Tool Drawing Provided Courtesy of Integrated Process Equip. Corp., 4717 E. Hilton Ave., Phoenix, AZ 85034.
28. Verteq Inc. 1241 E. Dyer Rd., Suite 100, Santa Ana, CA 92705.
29. Chen, L.-J., and C. C. Diao. In *A Novel In-Situ Thickness Measurement Method Using Pad Temperature Monitoring for CMP Technology*, *CMP–MIC Conference*, 241–8. Santa Clara, CA, 1996.
30. Koos, D. A., and S. Meikle. *Optical Endpoint Detection Methods in Semiconductor Planarizing Polishing Processes*. U.S. Patent 5,413,941, May 9, 1995.
31. Sandhu, G. S., L. D. Schultz, and T. T. Doan. *Method for Endpoint Detection During Chemical/Mechanical Planarization of Semiconductor Wafers*. U.S. Patent 5,036,015, July 30, 1991.

32. Yu, C. C., and G. S. Sandhu. *Chemical Mechanical Planarization (CMP) of a Semiconductor Wafer Using Acoustical Waves for In-Situ Endpoint Detection*. U.S. Patent 5,240,552, August 31, 1993.
33. Nova Measuring Instruments Inc., 1250 Oakmead Parkway, suite 210, Sunnyvale, CA 94088-3599.
34. Mirra Endpoint Tool Drawing Courtesy of Applied Materials, 3111 Coronado Dr., Santa Clara, CA 95054.
35. Jairath, R., M. Desai, M. Stell, R. Tolles, D. Scherber-Brewer. In *Consumables for the Chemical Mechanical Polishing (CMP) of Dielectrics and Conductors*, *Conference Proceedings Matter Research Society*, Vol. 337, 121–31. San Francisco, CA, 1994.
36. Rohm and Haas Corporation, 451 Bellevue Road, Newark, DE 19713.
37. Li, S., H. Banvillet, C. Augagneur, B. Miller, M. Nabot-Henaff, and K. Wooldridge. In *Evaluation of H_20_2, KIO_3 and $Fe(NO_3)_3$ Based W CMP Slurries for 8′ 0.35 μm CMOS Technology*, *CMP–MIC Proceedings*, 165. Santa Clara, CA, 1998.
38. Yen, B., F. Shau, W. Chiang, C. Huang, C. Yi. In *Evaluation of Two Types of Tungsten Slurries for Dual Damascene*, *CMP–MIC Proceedings*, 215. Santa Clara, CA, 1999.
39. Cadien, K., and D. Feller. *Slurries for Chemical Mechanical Polishing*. U.S. Patent 5,340,370, August 23, 1994.
40. Philipossian, A., M. Moinpour, and A. Oehler. In *An Overview of Current Issues and Future Trends in CMP Consumables*. 13. CMP–MIC Proceedings, Santa Clara, CA, 1996.
41. Neville M., D. Fluck, C. Hung, M. Lucarelli, and D. Scherber. *Chemical Mechanical Polishing Slurry for Metal Layers*. U.S. Patent 5,527,423, June 18, 1996.
42. Babu, S. V., M. Hariharaputhiran, S. Ramarajan, Y. Her, and M, Mayton. In *The Role of Particulate Properties in the Chemical–Mechanical Polishing of Copper*, *CMP–MIC Proceedings*, 121. 1998.
43. Cook, L., S. Loncki, and G. Brancaleoni. *Activated Polishing Compositions*. U.S. Patent 5,382,272, January 17, 1995.
44. Nojo, H., M. Kodera, and R. Nakata. In *Slurry Engineering for Self-Stopping, Dishing Free SiO_2–CMP, IEDM Proceedings*, 1996.
45. Maniar, D., and C. Yu. *Process for Polishing a Semiconductor Substrate*. U.S. Patent 5,525,191, June 11, 1996.
46. Kishi, S., R. Suzuki, A. Ohishi, and Y. Arimoto. In *Completely Planarized W Plugs Using MnO_2 CMP*, *IEDM Proceedings*, 465. 1995.
47. Kaufman, F. B., D. B. Thompson, R. E. Broadie, M. A. Jaso, W. L. Guthrie, D. J. Pearson, and M. B. Small. "Chemical Mechanical Polishing for Fabricating Patterned W Metal Features as Chip Interconnects." *J. Electrochem. Soc.* 138 (1991): 3460–5.
48. Kim, H. S., T. H. Lee, S. Y. Kim, and J. S. Choi. In *Performance Comparison of Consumable Materials for Tungsten Plug Chemical Mechanical Polishing Process*, *CMP–MIC Proceedings*, 457. Santa Clara, CA, 1998.
49. Skrovan, J., and K. Robinson. *Planarization Slurry Including a Dispersant and Method of Using Same*. U.S. Patent 5,827,781, October 27, 1998.
50. Luo, Q., D. Campbell, and S. V. Babu. "Stabilization of Alumina Slurries for Chemical–Mechanical Polishing of Copper." *Langmuir* 12 (1996): 3563.
51. Luo, Q., D. Campbell, and S. V. Babu. "Chemical–Mechanical Polishing of Copper in Alkaline Media." *Thin Solid Films* 311 (1997): 177.
52. Sasaki, Y., N. Hayasaka, H. Kaneko, H. Hirabayashi, and M. Higuchi. *Polishing Agent and Polishing Method Using the Same*. U.S. Patent 5,770,095, June 23, 1998.
53. Farkas, J., R. Jairath, M. Stell, and S. A. Tzeng. *Method of Using Additives with Silica Based Slurries to Enhance Selectivity in Metal CMP*. U.S. Patent 5,614,444, March 25, 1997.
54. Korman, R., and D. Capitano. "Distribution Systems for CMP: The New Challenge." *J. Electron. Mater.* 25 (1996): 1608.
55. Federau, M., and M. Maxim. *Cabot Corporation, Private Communication*.
56. Wilmer, J. *Control and Measurement of Critical CMP Slurry Parameters*, SEMI Education Series. Burlingame, CA: SEMICON West, 1998.
57. Pohl, M. C., and D. C. Griffiths. "The Importance of Particle Size to the Performance of the Abrasive Particles in the CMP Process." *J. Electron. Mater.* 25 (1996): 1612.

58. Gutmann, R., D. Price, J. Neyrick, C. Sainio, D. Permana, D. Duquette, and S. Murarka. In *CMP of Copper–Polymer Interconnect Structures*, *CMP–MIC Proceedings*, 251. Santa Clara, CA, 1998.
59. Grover, G. S., H. Liang, S. Ganeshkumar, and W. Fortino. "Effect of Slurry Viscosity Modification on Oxide and Tungsten CMP." *Wear* 214 (1998): 10.
60. Liu, B., S. Yoo, H. Chung, and S. Chae. *Comparative Particle Size Measurement of CMP Slurries by Instrumental and Filtration Methods*, CMP Technology for ULSI Interconnection. Burlingame, CA: SEMICON West, 1998.
61. Bare, J. *Comparison of Vacuum-Pressure vs. Pump Dispense Engines for CMP Slurry Distribution*, Semi Workshop on Contamination in Liquid Chemical Distribution Systems. Burlingame, CA: SEMICON West, 1998.
62. Bare, J., and T. Lemke. "Monitoring Slurry Stability to Reduce Process Variability." *Micro* September (1997): 53.
63. Corlett, G. "Can CMP Waste Ever Be Environmentally Friendly?" *J. Adv. Appl. Contam. Control.* 1-11 (1998): 19.
64. Mendocino, L., and P. T. Brown. "The Environmental, Health and Safety Side of Copper Metallization." *Semicond. Int.* June (1998): 105.
65. Brown, N. "Preparation of Ultrasmooth Surfaces." *Ann. Rev. Mater. Sci.* 16 (1986): 371–88.
66. Brown, N., P. Baker, and R. Maney. "Optical Polishing of Metals." *Proc. SPIE* 306 (1981): 42–57.
67. Brown, N., and L. Cook, Paper TuB-A4, In *Technical Digest*, *Topical Meeting on the Science of Polishing*, Optical Society of America, April 1984.
68. Liu, C. W., B. T. Dai, and C. F. Yeh. "Characterization Pf the Chemical–Mechanical Polishing Process Based on Nanoindentaation Measrement of Dielectric Films." *J. Electrochem. Soc.* 142 (1995): 3098–104.
69. Liu, C. W., B. T. Dai, W. T. Tseng, and C. F. Yeh. "Modeling of the Wear Mechanism during Chemical–Mechanical Polishing." *J. Electrochem. Soc.* 143 (1996): 716.
70. Kallingal, C. G., M. Tomozawa, and S. P. Muraka. "Substrate Effects on Hardness and Polishing of SiO_2 Thin Films." *J. Electrochem. Soc.* 145 (1998): 1790–4.
71. Runnels, S. R., and L. M. Eyman. "Tribology Analysis of Chemical–Mechanical Polishing." *J. Electrochem. Soc.* 141 (1994): 1698–701.
72. Tseng, W. T., and Y. L. Wang. "Re-Examination of Pressure and Speed Dependance of Removal Rates during Chemical–Mechanical Polishing Processes." *J. Electrochem. Soc.* 144 (1997): L15–7.
73. Shi, F. G., and B. Zhao. "Modeling of Chemical–Mechanical Polishing with Soft Pads." *Appl. Phys. A* 67 (1998): 249–52.
74. Shi, F. G., B. Zhao, and S. Q. Wang. In *A New Theory for CMP with Soft Pads*, *Conference Proceedings of IITC*, 73–5. San Francisco, CA, 1998.
75. Zhao, B., and F. G. Shi. "Chemical Mechanical Polishing: Threshold Pressure and Mechanism." *Electrochem. Solid State Lett.* 2 (1999): 145–7.
76. van Kranenburg, H., H. D. van Corbach, P. H. Woerlee, and M. Lohrmeier. "W-CMP for Sub-Micron Inverse Metalization." *Microelectron. Eng.* 33 (1997): 241–8.
77. Hansen, D. A., et al. In *Proceeding CMP for ULSI Multilevel Interconnection Conference*, 227. 1997.
78. Stavreva, Z., D. Zeidler, M. Plotner, G. Grasshoff, and K. Drescher. "Chemical–Mechanical Polishing of Copper for Interconnect Formation." *Microelectron. Eng.* 33 (1997): 249–57.
79. Baker, A. R. *Conf. Proc. Electrochem. Soc.* 96-22 (1996): 228.
80. Wang, J. L., K. Holland, T. Bibby, S. Beaudoin, and T. Cale. "Van Mises Stress in Chemical–Mechanical Polishing Processes." *J. Electrochem. Soc.* 144 (1997): 1121–7.
81. Wijekoon, K., R. Lin, B. Fishkin, S. Yang, F. Redeker, G. Amico, and S. Nanjangud. "Tungsten CMP Process Developed." *Solid State Technol.* 4 (1998): 53–6.
82. Burke, P. A. In *Conference Proceedings CMP–MIC*, 379. Santa Clara, CA, 1991.
83. Warnock, J. "A Two-Dimensional Process Model for Chemimechanical Polish Planarizaton." *J. Electrochem. Soc.* 138 (1991): 2398–402.
84. Runnels, S. R. "Feature-Scale Fluid-Based Erosion Modeling for Chemical–Mechanical Polishing." *J. Electrochem. Soc.* 141 (1994): 1900–4.

85. Runnels, S. R., I. Kim, J. Schleuter, C. Karlsrud, and M. Desai. "Modeling Tool for Chemical–Mechanical Polishing Design and Evaluation." *IEEE Trans. Semicond. Manufact.* 11 (1998): 501–10.
86. Yu, T. K., C. C. Yu, M. Orlowski. In *Statistical Polishing Pad Model for Chemical–Mechanical Polishing*, *Technical Digest IEDM*, 865–8. Washington, DC, 1993.
87. Chekina, O. G., L. M. Keer, and H. Liang. "Wear-Contact Problems and Modeling of Chemical Mechanical Polishing." *J. Electrochem. Soc.* 145 (1998): 2100–6.
88. Tseng, E., C. Yi, and H. C. Chen. In *Conference Proceedings CMP–MIC*, 258. Santa Clara, CA, 1997.
89. Grillaert, J., M. Meuris, N. Heylen, K. Devriendt, E. Vrancken, and M. Heyns. In *Conference Proceedings CMP–MIC*, 79. Santa Clara, CA, 1998.
90. Stine, B. E., D. S. Boning, and J. E. Chung. "Analysis and Decomposition of Spatial Variation in Integrated Circuits Processes and Devices." *IEEE Trans. Semicond. Manufact.* 10 (1997): 24–41.
91. Chang, E., B. Stine, T. Maung, R. Divecha, D. Boning, J. Chung, K. Chang, et al. *Using a Statistical Metrology Framework to Identify Systematic and Random Sources of Die-Level and Wafer-Level ILD Thickness Variations in CMP Processes*, *Technical Digest IEDM*, 499–502, Piscataway, NJ, 1995.
92. Stine, B. E., D. O. Ouma, R. R. Divecha, D. S. Boning, J. E. Chung, D. L. Hetherington, C. R. Harwood, O. S. Nakagawa, and S. Y. Oh. "Rapid Characterization and Modeling of Pattern-Dependant Variation in Chemical–Mechanical Polishing." *IEEE Trans. Semicond. Manufact.* 11 (1998): 129–40.
93. Ouma, D., D. Boning, J. Chung, G. Shinn, L. Olsen, and J. Clark. In *An Integrated Characterization and Modeling Methodology for CMP Dielectric Planarization. Conference Proceedings of International Interconnect Technology Conference*, 67–9. San Francisco, CA, 1998.
94. Ouma, D., C. Oji, D. Boning, J. Chung, D. Hetherington, and P. Merkle. In *Conference Proceedings CMP–MIC*, 20. Santa Clara, CA, 1998.
95. Fang, S. J., T. H. Smith, G. B. Shinn, J. A. Stefani, and D. S. Boning. In *Advanced Process Control in Dielectric Chemical Mechanical Polishing (CMP). Conference Proceedings CMP–MIC*, 367–74. Santa Clara, CA, 1999.
96. Smith, T. H., S. J. Fang, D. S. Boning, G. B. Shinn, and J. A. Stefani. In *Conference Proceedings CMP–MIC*, 97. Santa Clara, CA, 1999.
97. Elbel, N., B. Neureither, B. Ebersberger, and P. Lahnor. "Tungsten Chemical Mechanical Polishing." *J. Electrochem. Soc.* 145, no. 5 (1998): 1659–64.
98. Steigerwald, J. M., R. Zirpoli, S. P. Murarka, D. Price, and R. J. Gutmann. "Pattern Geometry Effects in the Chemical–Mechanical Polishing of Inlaid Copper Structures." *J. Electrochem. Soc.* 141 (1994): 2842.
99. Stine, B. E., R. Vallishayee. In *On the Impact of Dishing in Metal CMP Processes on Circuit Performance*, *IEEE International Workshop Statistical Metrology*, 64–7. 1998.
100. Cook, L. "Chemical Processes in Glass Polishing." *J. Non-Cryst. Solids* 120 (1990): 152–71.
101. Nogami, M., and M. Tomozawa. "Diffusion of Water in High Silica Glasses as Low Temperature." *Phys. Chem. Glass.* 25 (1984): 82–5.
102. King, J. A., Ed. In *Materials Handbook for Hybrid Microelectronics*, Artech House, Inc. 1988.
103. Pourbaix, M. Atlas of Electrochemical Equilibria in Aqueous Solutions. Houston, TX: National Association of Corrosion Engineering, 1974.
104. Osseo-Asare, K., M. Anik, and J. DeSimone. "Chemical Mechanical Polishing of Tungsten: Effect of Tungstate Ion on the Electrochemical Behavior of Tungsten." *Electrochem. Solid State Lett.* 2 (1999): 143–4.
105. Johnson, H. E., and J. Leja. *J. Electrochem. Soc.* 112 (1965): 638.
106. Carpio, R., J. Farkas, and R. Jairath. "Initial Study on Copper CMP Slurry Chemistries." *Thin Solid Films* 266 (1995): 238–44.
107. Luo, Q., S. V. Babu, In *Conference Proceedings CMP–MIC*, 83. Santa Clara, CA, 1997.
108. Landis, H., P. Burke, W. Cote, W. Hill, C. Hoffman, C. Kaanta, C. Koburger, W. Lange, M. Leach, and S. Luce. "Chemical–Mechanical Polishing in CMOS Integrated Circuits." *Thin Solid Films* 220 (1992): 1–7.
109. Yu, C., C. Fazan, V. Matthews, and T. Doan. "Dishing Effects in a Chemical–Mechanical Polishing Planarization Process for Advanced Trench Isolation." *Appl. Phys. Lett.* 61 (1992): 1344.

110. Davari, B., C. Koburger, R. Schulz, J. Warnock, T. Furukawa, M. Jost, Y. Taur, et al. "A New Planarization Technique Using a Combination of RIE and Chemical Mechanical Polish (CMP)." *IEDM Tech. Dig.* (1989): 61–4.
111. Daubenspeck, T., J. DeBrosse, C. Koburger, M. Armacost, and J. Abernathy. "Planarization of ULSI Topography Over Variable Pattern Densities." *J. Electrochem. Soc.* 138 (1991): 506–9.
112. Cooperman, S., A. Nasar, and G. Grula. "Optimization of a Shallow Trench Isolation Process for Improved Planarization." *J. Electrochem. Soc.* 142 (1995): 3180–5.
113. Jouty, M., M. Rivoire, and T. Detzel. In *The Effect of Feature Size and Counter Mask on Oxide Removal Rate in Shallow Trench Isolation*, *Conference Proceedings CMP–MIC*, 329–32. Santa Clara, CA, 1999.
114. Lao, P., T. Tsai, S. Lin, C. Lee, E. Hsu, H. Wu, H. Chen, and L. Liu. In *Characterization of Selective CMP, Dummy Pattern and Reverse Mask Approaches in STI Planarization Process*, *Conference Proceedings CMP–MIC*, 333–5. Santa Clara, CA, 1999.
115. Beyer, K., J. Makris, E. Mendel, K. Numtay, S. Ogura, J. Riseman, and N. Rovedo. *Method for Removing Protuberances at the Surface of a Semiconductor Wafer Using a Chem–Mech Polishing Technique*. U.S. Patent 4,671,851. June 9, 1987.
116. Boyd, J., and J. Ellul. "A One-Step Shallow Trench Global Planarization Process Using Chemical Mechanical Polishing." *Electrochem. Soc. Proc.* (1995): 290–301.
117. Lin, M., C. Y. Chang, D. Liao, B. Wang, and A. Henderson. In *Improved STI CMP Technology for Micro-Scratch Issue*. *Conference Proceeding CMP–MIC*, 322–6. Santa Clara, CA, 1999.
118. Kim, C., S. Lee, J. Kim, M. Kim, S. Hong, S. Hah, U. Chung, and M. Lee. In *Selective CMP (Chemical–Mechanical Polishing) Using BN (Boron Nitride) Films to Achieve Global Planarization*. *Conference Proceeding VMIC*, 401–6. 1996.
119. Grillaert, J., N. Heylen, E. Vranchen, G. Badenes, R. Rooyackers, M. Meuris, and M. Heynes. In *A Novel Approach for the Elimination of Pattern Density Dependence of CMP for Shallow Trench Isolation*, *Conference Proceedings CMP–MIC*, 313. Santa Clara, CA, 1998.
120. Choi, K. S., S. I. Lee, C. I. Kim, C. W. Nam, S. D. Kim, and C. T. Kim. In *Application of Ceria-Based High Selectivity Slurry to STI CMP for sub-0.18 Micron CMOS Technologies*, *Conference Proceeding CMP–MIC*, 307–13. Santa Clara, CA, 1999.
121. Stine, B., D. Boning, J. Chung, L. Camilletti, F. Kruppa, E. Equi, W. Loh, et al. "The Physical and Electrical Effects of Metal-Fill Patterning Practices for Oxide Chemical–Mechanical Polishing Processes." *IEEE Trans. Electron. Dev.* 45 (1998): 665–79.
122. Camilletti, L. In *Implementation of CMP-Based Design Rules and Patterning Practices*, *IEEE/SEMI Advance Semiconductor Manufacturing Conference*, 2–4. 1995.
123. Ouma, D. Modeling of Chemical Mechanical Polishing for Dielectric Planarization. Ph.D. thesis. MIT, Cambridge, MA, 1998.
124. Ireland, P. J. "High Aspect Ratio Contacts: A Review of the Current Tungsten Plug Process." *Thin Solid Films* 304 (1997): 1–12.
125. Brusic, V., G. S. Frankel, C.-K. Hu, M. M. Plechaty, and G. C. Scwartz. "Corrosion and Protection of Thin-Line Conductors in VLSI Structures." *IBM J. Res. Dev.* 37, no. 2 (1993): 173.
126. Brusic, V., M. A. Frisch, B. N. Eldridge, F. P. Novak, F. B. Kaufman, B. M. Rush, and G. S. Frankel. "Copper Corrosion with and without Inhibitors." *J. Electrochem. Soc.* 138 (1991): 2253–9.
127. Gangulee, A. "Structure of Electroplated and Vapor-Deposited Copper Films. II. Effects of Annealing." *J. Appl. Phys.* 43 (1972): 3943.
128. Vanasupa, L., D. Pinck, Y.-C. Joo, T. Nogami, S. Pramanick, S. Lopatin, and K. Yang. "The Impact of Linewidth and Line Density on the Texture of Electroplated Cu in Damascene-Fabricated Lines." *Electrochem. Solid State Lett.* 2, no. 6 (1999): 329.
129. Sethuraman, A. R., J.-F. Wang, and L. M. Cook. "Copper vs Aluminum: A Planarization Perspective." *Semicond. Int.* June (1996): 177.
130. Wang, J.-F., A. R. Sethuraman, L. M. Cook, R. C. Kistler, and G. P. Schwartz. "Chemical–Mechanical Polishing of Dual Damascene Aluminum Interconnect Structures." *Semicond. Int.* October (1995): 117–22.

131. Wang, Y. L., J. Wu, C. W. Liu, T. C. Wang, and J. Dun. "Material Characteristics, Chemical–Mechanical Polishing of Aluminum Alloy Thin Films." *Thin Solid Films* 332 (1998): 397–403.
132. Krussel, W. C., J. M. deLarios, and J. Zhang. "Mechanical Brush Scrubbing for Post-CMP Clean." *Solid State Technol.* 38, no. 6 (1995): 109.
133. Roy, S. R., I. Ali, G. Shinn, N. Furusawa, R. Shah, S. Peterman, K. Witt, S. Eastman, and P. Kumar. "Post Chemical–Mechanical Planarization Cleanup Process for Interlayer Dielectric Films." *J. Electrochem. Soc.* 142 (1995): 216–26.
134. Hymes, D. J., and I. J. Malik. "Using Double-Sided Scrubbing Systems for Multiple General Fab Applications." *Micro* 14 (1996): 7.
135. Fraser, B., M. B. Olsen, T. Phan, B. Morrison. *Conference Proceeding Electrochemistry Society*, 634. 1997.
136. Takenaka, N., Y. Satoh, A. Ishihama, K. Sakiyama. In *Post CMP Cleaning Using Ice Scrubber Cleaning*, *Conference Proceeding Mater Research Society*. Vol. 386, 121–5. San Francisco, CA, 1995.
137. Menon, V. B., and R. P. Donovan. In *Handbook of Semiconductor Wafer Cleaning Technology*, edited by W. Kern, 379. Park Ridge, NJ: Noyes Publications, 1993.
138. Shwartzman, S., A. Mayer, and W. Kern. "Megasonic Particle Removal from Solid-State Wafers." *RCA Rev.* 46-3 (1985): 81.
139. Shen, J. J., W. D. Costas, L. M. Cook, and J. Farber. "Effects of Post Chemical Mechanical Planarization Buffing on Defect Density of Tungsten and Oxide Wafers." *J. Electrochem. Soc.* 145 (1998): 4240–3.
140. Ali, I., S. R. Roy, and G. B. Shinn. "Investigating the Effect of Secondary Platen Pressure on Post-Chemical–Mechanical Planarization Cleaning." *Microcontamination* 12-10 (1994): 45–50.
141. Cohen, S. A., M. A. Jaso, and A. A. Bright. "Electrical Properties of Chemical–Mechanical Polished Tetraethyl Orthosilicate Films with and Without Capping Layers." *J. Electrochem. Soc.* 139 (1992): 3572–4.
142. Murarka, S., S. H. Ko, M. Tomozawa, P. J. Ding, W. A. Lanford. In *Surface Damage in SiO_2 Caused by Chemical Mechanical Polishing on IC-60 Pads*, *Conference Proceeding Material Research Society*. Vol. 377, 157–165. 1994.
143. Wallace, W. E., W. L. Wu, and R. A. Carpio. "Chemical–Mechanical Polishing of SiO_2 Thin Films Studied by X-Ray Reflectivity." *Thin Solid Films* 280 (1996): 37–42.
144. Trogolo, J. A., and K. Rajan. "Near Surface Modification of Silica Structure Induced by Chemical/Mechanical Polishing." *J. Mat. Sci.* 29 (1994): 4548–54.
145. Steigerwald, J. M., S. P. Murarka, J. Ho, R. J. Gutman, and D. J. Duquette. "Mechanisms of Copper Removal during Chemical Mechanical Polishing." *J. Vac. Sci. Technol. B* 13, no. 6 (1995): 2215–8.
146. Osseo-Asare, K., and K. K. Mishra. "Solution Chemical Constraints in the Chemical–Mechanical Polishing of Copper: Aqueous Stability Diagrams for the Cu-H_2O and Cu-NH_3-H_2O Systems." *J. Electron. Mater.* 25 (1996): 1599–607.
147. Hymes, D., H. Li, E. Zhao, and J. de Larios. "The Challenges of the Copper CMP Clean." *Semiconduct. Intl* June (1998): 117–22.
148. Feng, Y., W.-K. Teo, K.-S. Siow, Z. Gao, K.-L. Tan, and A.-K. Hsieh. "Corrosion Protection ofCopper by a Self-Assembled Monolayer of Alkanethiol." *J. Electrochem. Soc.* 144 (1997): 55–64.
149. Wang, M. T., M. S. Tsai, C. Liu, W. T. Tseng, T. C. Chang, L. J. Chen, and M. C. Chen. "Effects of Corrosion Environments on the Surface Finishing of Copper Chemical Mechanical Polishing." *Thin Solid Films* 308–309 (1997): 520–2.
150. Tromans, D. "Aqueous Potential-pH Equilibria in Copper-Benzotriazole Systems." *J. Electrochem. Soc.* 145 (1998): L43–5.
151. Osaka, T., and T. Hattori. "Influence of Initial Wafer Cleanliness on Metal Removal Efficiency in Immersion SC-1 Cleaning: Limitation of Immersion-Type Wet Cleaning, IEEE Trans." *Semicond. Manufcat.* 11, no. 1 (1998): 20–4.
152. de Larios, J. M., M. Ravkin, D. L. Hetherington, and J. J. Doyle. "Post CMP Cleaning for Oxide and Tungsten Applications." *Semicond. Int.* 121, (1996).
153. Malik, I. J., J. Zhang, A. J. Jensen, J. J. Farber, W. C. Krusell, S. Raghavan, C. Rajhunath. In *Post-CMP Cleaning of W and SiO_2: A Model Review*, *Material Research Society Symposium*. 386. 1995.

18

Optical Lithography

18.1 Introduction **18**-1
18.2 Patterning Basics **18**-2
Pattern Generators • Pattern Replicators • Imaging Basics
18.3 Optics for Manufacturing **18**-7
Aberrations
18.4 Exposure Tool System Considerations **18**-10
Alignment and Overlay • Throughput
18.5 Resolution Enhancement Techniques **18**-19
Phase Shift Masks • Modified Illumination • Optical Proximity Effect
18.6 Manufacturing Considerations **18**-32
Limits to Optical Lithography • Process Latitude • Mask Error Factor • Mix-and-Match Lithography • Error Budgets
18.7 Recent Advances in Optical Lithography **18**-38
Immersion Lithography • Polarization
18.8 Patterning Roadmaps **18**-45
18.9 Summary **18**-46
References **18**-47

Gene E. Fuller
Strategic Lithography Services

18.1 Introduction

A fundamental requirement for almost all useful semiconductor devices is the definition of patterned elements. The overwhelming technology choice for performing this patterning since the very inception of semiconductor manufacturing has been optical lithography. Up to the early 1970s, most lithography was performed as a contact or close proximity printing process in which blue and near UV light was passed through a photomask directly onto a photoresist-coated semiconductor substrate [1]. This simple-appearing shadow imaging process has been described in many research publications and handbooks [2,3], and will not be further discussed in this chapter. There is very little remaining use of contact or proximity printing in the very large scale integration (VLSI) manufacturing world.

In the 1970s, the first widespread use of projection printing for semiconductor manufacturing was fostered by a very well-accepted family of tools from Perkin-Elmer, the so-called Micralign projection aligners. These tools for the first time allowed higher performance pattern definition by scanning and imaging only a fractional area of the wafer at any instant. Both the optical resolution performance and the pattern overlay performance were significantly enhanced over the performance possible with whole-wafer patterning tools. This became particularly important as the typical wafer size in manufacturing reached 3 in. in diameter.

Beginning in the late 1970s and early 1980s, a new class of projection exposure tools, typically known as steppers, was introduced [4]. For the first time, the pattern definition imaging on the semiconductor wafers was performed one chip at a time in a step-and-repeat fashion. This had profound implications on the requirements for the photomask as well as the precision mechanical movements needed to accurately overlay a new pattern to underlying patterns already on the wafer substrate. Most stepper systems employed a reduction projection lens to ease the fabrication difficulty of the photomask and to improve the overall precision and accuracy of the overlay of patterns on the wafer. The photomask used for step-and-repeat lithography is often called a "reticle." Both terms will be used throughout this chapter.

This general class of step-and-repeat exposure tools, namely optical steppers, has served as the dominant patterning technology for the past 25 years, and is expected to last for a number of years into the future. Key features of the stepper tools include

- Limited imaging area at any instant, allowing highest possible resolution performance for a given complexity of imaging lens.
- Pattern overlay can be performed independently for these small field areas across the wafer. This allows correction for either wafer-induced distortion or lens-induced distortion.
- The small field area is exposed identically over the entire wafer. This allows the reticle to be manufactured at a scale larger than 1:1, greatly increasing the effective precision and quality of placing pattern objects on the reticle.

Following the steppers, with significant commercial introduction in the mid-1990s, a combination of the earlier scanning approach with the step-and-repeat approach was created. In this instance, each step-and-repeat field is imaged in a scanning fashion rather than all at once in a single flash. The first such tool in widespread use was the Micrascan introduced by Silicon Valley Group. The step-and-scan approach, as it is often called, spread rapidly throughout the lithography tool industry, and almost all of the new deep UV exposure tools now being delivered are scanning tools.

18.2 Patterning Basics

Microlithography is the process of defining useful shapes on the surface of a semiconductor wafer. Typically, this consists of a patterned exposure into some sort of photosensitive material already deposited on the wafer. A variety of processes that directly pattern the wafer are possible, such as directly writing on the wafer with an electron beam, but at this time none are in use for high-volume semiconductor manufacturing.

In the cases of interest in this chapter, the photosensitive material is an organic material known as photoresist that consists of a polymer base resin with additional components to provide photosensitivity and other properties. There is an extensive discussion of photoresists in a separate chapter in this volume.

The patterned exposure of the desired wafer image can be performed by a number of techniques, as illustrated in Figure 18.1. The two major categories of patterning systems are pattern generators and pattern replicators. Virtually all semiconductor microlithography is performed using pattern replicators, but it is instructive to consider the differences.

18.2.1 Pattern Generators

The definition of a pattern generator is an exposure tool that accepts pattern input data from a database and directly creates the physical image on the wafer. These tools are used extensively to create photomasks and reticles for pattern replicators, and they are also used in limited volume for direct patterning on semiconductor wafers.

Pattern generators use either charged particles or photons to create the image. Photomask fabrication is performed with electron beam and photon beam tools, while repair of mask pattern defects is performed with ion beam and photon beam tools.

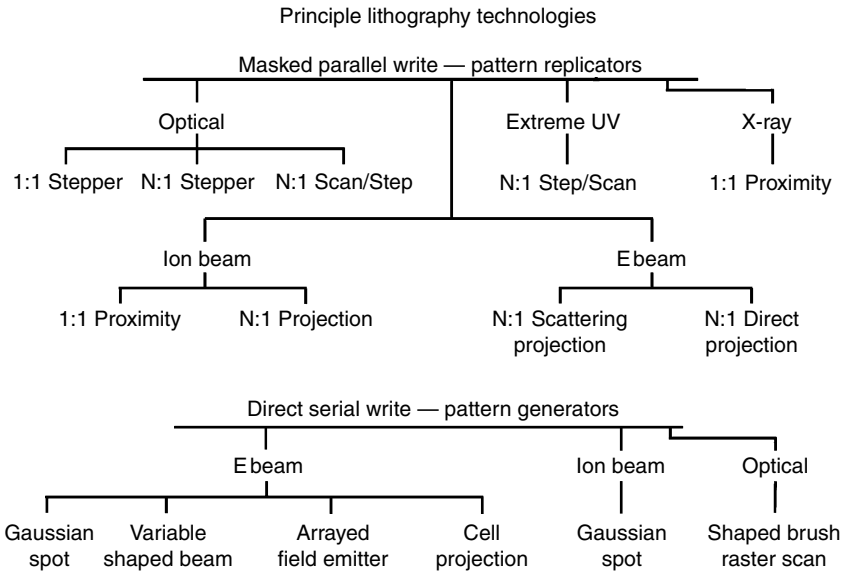


FIGURE 18.1 Lithography exposure tool options.

The key disadvantage in using pattern generators for general purpose high-volume lithography is the slow imaging on the wafer. The required patterning density is on the order of $10^{13}/cm^2$ discrete pixels for 65 nm circuits being developed and introduced as this is written. The achievable speeds for electron or photon pattern generators are less than 10^{10} pixels/s, leading to imaging rates of no more than one 200 mm wafer per hour. The general rule for economical semiconductor production requires a wafer patterning tool to process on the order of 100 wafers per hour or higher, and existing direct pattern generators simply cannot come close to achieving this speed.

18.2.2 Pattern Replicators

The solution to the throughput limitation in pattern generators is to create a master pattern image in the form of a photomask or reticle and then replicate the pattern in a massively parallel fashion onto the wafer. A variety of image transfer techniques can be used, including photons and charged particles. The most common pattern transfer agent is a well-conditioned beam of monochromatic photons in the wavelength regime from 193 to 436 nm.

The first practical step-and-repeat imaging tools used the so-called g-line of mercury, at 436 nm wavelength. Second generation stepper tools used the i-line of mercury, at 365 nm. More recently, as the need for higher resolution drove the requirement for wavelength down, mercury arc lamps were replaced by excimer lasers. The lasers provide both very high intensity and very narrow bandwidth. The most common laser in use today is the KrF laser at 248 nm wavelength. The most advanced exposure tools today in the semiconductor manufacturing scene employ an even shorter wavelength, 193 nm, from the ArF excimer laser.

The relationship of wavelength to image resolution and depth of focus in a projection optical system has been understood for more than 100 years. The simple relationship defined by the so-called Rayleigh equations is shown in Figure 18.2.

The simple expressions for resolution and depth of focus have been widely used for many years.

$$R = k_1 \frac{\lambda}{\mathrm{NA}} \tag{18.1}$$

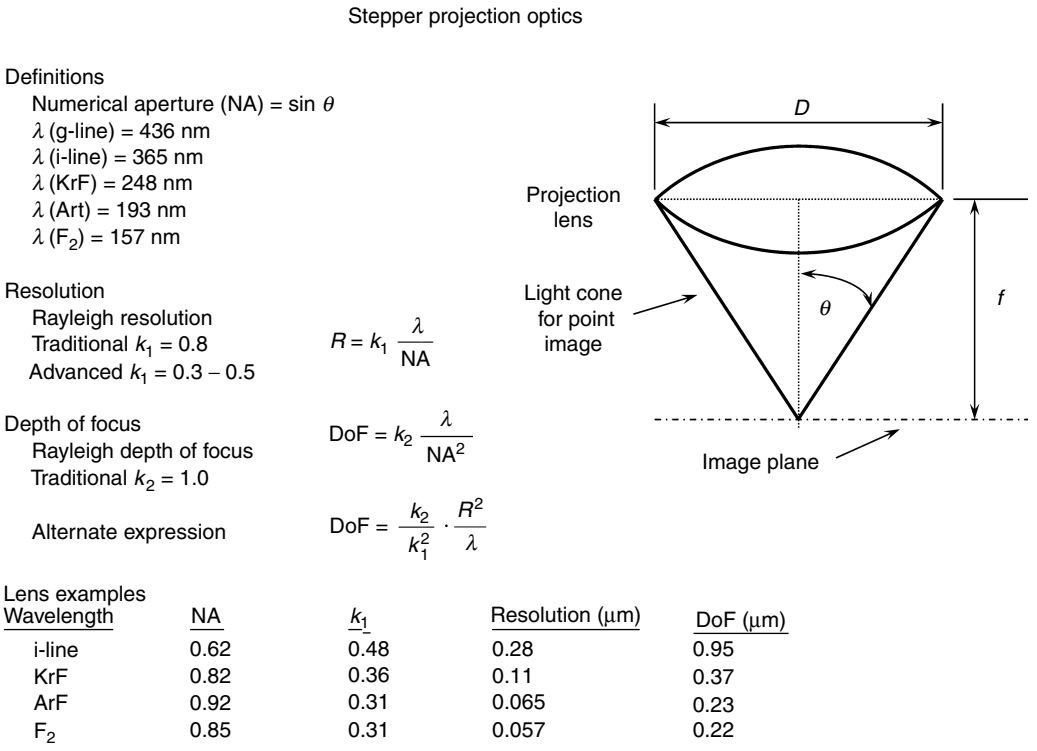


FIGURE 18.2 Basic relationships in projection optical systems.

$$\mathrm{DoF} = k_2 \frac{\lambda}{\mathrm{NA}^2} \tag{18.2}$$

These expressions provide useful guidelines when the maximum angle of the light rays is relatively small, or in other words when the numerical aperture (NA) is small. In this case, the standard paraxial approximation of setting sin θ equal to θ is reasonable. For larger angles and NA values, the error in using this simple approximation grows unacceptably large. Lin has discussed this issue in some detail [5]. He has proposed that the original Rayleigh equations be replaced with a new pair of equations.

$$R = k_1 \frac{\lambda}{\sin\theta} \tag{18.3}$$

$$\mathrm{DoF} = k_3 \frac{\lambda}{\sin^2(\theta/2)} \tag{18.4}$$

18.2.3 Imaging Basics

Practical lithography is based on replicating the pattern defined on a photomask into resist-covered wafers. In an ideal case with no degradation at all in the imaging process, a simple copy of the mask pattern would result, as shown in Figure 18.3a. However, in a projection imaging process, the imaging is always subject to degradation from diffraction and imperfections in the projection system. An example of the image from a diffraction-limited projection system is shown in Figure 18.3b.

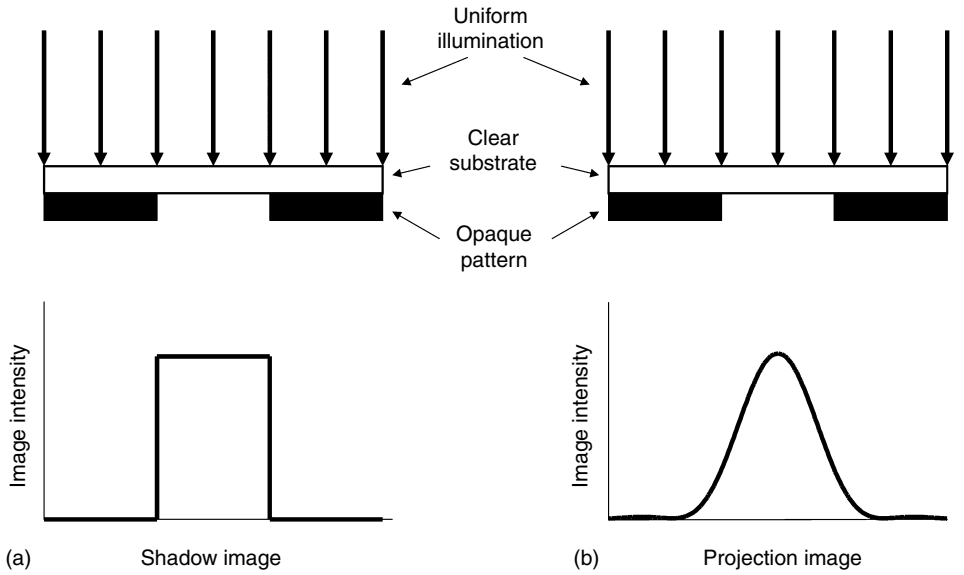


FIGURE 18.3 Basic imaging characteristics. (a) ideal shadow imaging and (b) diffraction-broadened projection imaging.

The fundamental optical properties behind the functional form of the image shown in Figure 18.3b are well understood, and have been thoroughly studied for nearly 200 years. The spreading of the image profile results from the wave nature of light, and it is this property that effectively limits the resolution capability of optical imaging systems. In an imaging lens system with a circular aperture of radius r and imaging distance z, the image intensity resulting from a point source can be described by an expression containing a first-order Bessel function,

$$I(\rho) = I_0 \left(2 \frac{J_1(x)}{x} \right)^2, \tag{18.5}$$

where $x = \rho/z \times 2\pi/\lambda \times r$ and ρ is the distance in the image plane from the geometrical image point. Through further consideration of the imaging characteristics of projection optical systems, the expression for x can be further simplified to $x = 2\pi \times \rho \times \mathrm{NA}/\lambda$, where NA is the numerical aperture of the projection lens. A simplified description of NA is given in Figure 18.2.

This light intensity distribution is known as the Airy pattern, after G. B. Airy, who first derived it in 1835 [6]. In addition to the general shape of the curve, shown in Figure 18.4, the first zero value and the first maximum value are of interest. These occur, respectively, at coordinates 3.832 and 5.136, leading to an image intensity minimum at $x = 0.61 \times 2\pi$ and an intensity maximum at $x = 0.82 \times 2\pi$. These special values will be further discussed below.

Resolution is defined as the ability to distinguish separate components of an object or a group of objects. The resolution capability of astronomical telescopes was studied in detail by Lord Rayleigh in the 19th century [7]. He defined the limit of resolution for a telescope as the angular separation between two stars when the peak of the Airy intensity pattern from one star coincided with the first minimum of the Airy intensity pattern for the other star. This leads to the well-known Rayleigh condition for angular resolution,

$$\sin\theta = 0.61 \frac{\lambda}{r}, \tag{18.6}$$

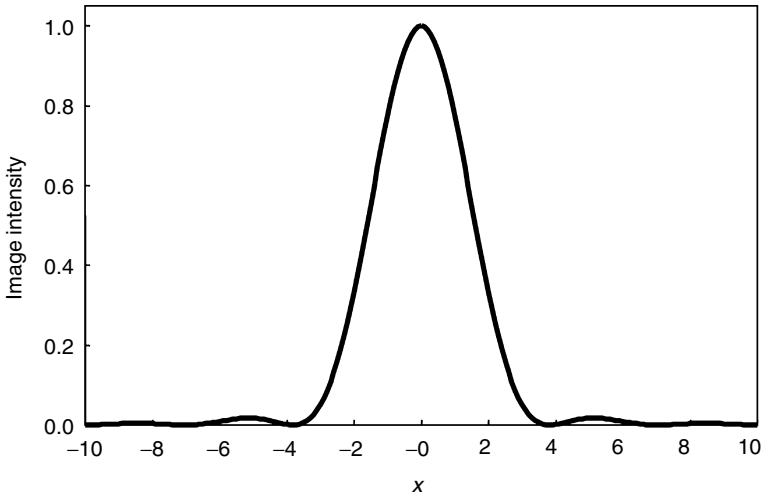


FIGURE 18.4 Airy pattern. Light intensity distribution from a point source projected through a circular imaging lens. The horizontal variable, x, is defined in the text.

where r is the radius of the imaging objective aperture. A sketch of the Rayleigh resolution condition is shown in Figure 18.5. Note that the intensity at the midpoint between the image peaks is reduced to about 78% of the peak intensity, which provides discernable separation, but not with high contrast between the bright and dark regions.

While the analogy of astronomical imaging to photolithography is not completely quantitative some key observations can be made. There is a limit to resolution for any given projection optical system, and it is not possible to resolve arbitrarily small or closely spaced features. It is also apparent that the resolution can be improved by using a smaller wavelength of the exposure light, and the resolution can be improved by making the projection system aperture larger.

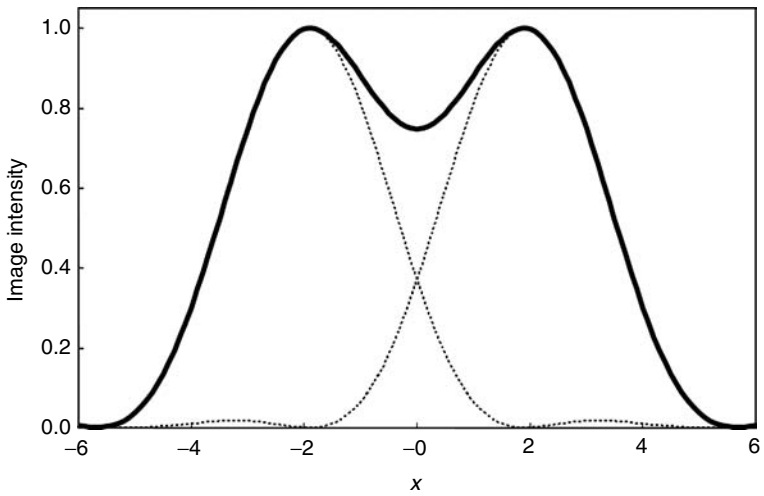


FIGURE 18.5 Rayleigh criterion for resolution of two point images. Horizontal scale is the same as in Figure 18.4.

In practical lithography, the Rayleigh condition is typically restructured into the Rayleigh equation described previously,

$$\text{Resolution} = k_1 \frac{\lambda}{\text{NA}}, \tag{18.7}$$

where NA is the numerical aperture of the projection system and k_1 is a constant on the order of 0.3–0.8. There is no rigorous optical definition for the constant k_1, and it is generally used as a qualitative descriptor of the overall lithography process capability. This common description of resolution capability is closely related to the Airy pattern described above. In particular, the first minimum of the Airy pattern occurs at $\rho = 0.61 \times \lambda/\text{NA}$ and the first maximum occurs at $\rho = 0.82 \times \lambda/\text{NA}$. The qualitative agreement with the usual range of k_1 constants is apparent.

18.3 Optics for Manufacturing

Virtually all of the exposure tools used today in manufacturing of silicon wafers rely on projection optical systems to form the patterns of the wafers. The requirements for these projection optics have always been quite stringent, but with the rapid progression in the worldwide semiconductor technology roadmap the optical requirements today are even more challenging.

The various technical needs for imaging tools can be reduced to a single principle:

All pattern feature edges must be placed in the correct location on the wafer, within the tolerances established for the specific product layer.

While this statement appears simple, it incorporates critical dimension (CD) control, alignment and overlay, lens aberrations, reticle errors, resist processing, and many other details. From the optics perspective, it requires high-resolution imaging with very low placement errors and uniform performance across the imaging field.

Another important constraint placed on all imaging system is related to cost. Certainly, the component cost and manufacturing cost are key elements of total cost, but in addition the cost includes the impact of defects added to the wafer, throughput capability, operating cost, and so on.

Two major classes of projection optics have evolved to meet the challenge of high-quality image formation. The most widely used optical format over the past 20 years has been the fully refractive imaging lens, working at a monochromatic wavelength. Such lenses have become large, complex, and expensive. Up to 30 separate lens elements are used, and the remaining aberrations are reduced to less than 1% of the wavelength. This format of imaging system is currently used by all of the major exposure tool suppliers in the world.

The second class of projection optics suitable for lithography use is the catadioptric form, which uses mirrors as primary imaging formation components in addition to refractive lenses. One key feature of such optics is that the allowable wavelength spread in the imaging light can be much broader than in the purely refractive systems. The catadioptric format most prominently used for advanced lithography was offered by SVG lithography in the Micrascan series of tools. A sketch of this imaging system is shown in Figure 18.6. These systems are no longer developed and sold as new. However, as discussed later in Section 18.7 the use of catadioptric projection optics is expected to return as high NA immersion lithography tools are developed.

18.3.1 Aberrations

An ideal lens would project a point object on the reticle into a small point at the geometrically correct location on the wafer, subject only to the diffraction broadening discussed in Section 18.2. In real optical

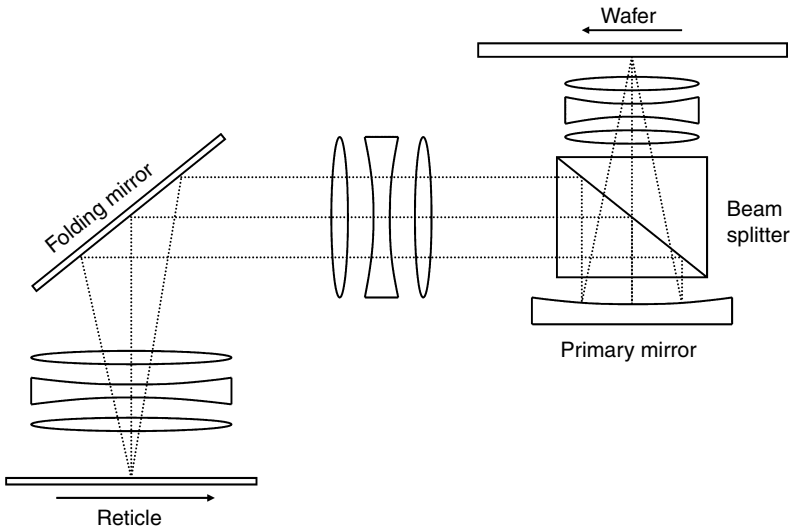


FIGURE 18.6 Schematic view of catadioptric projection system similar to that used in SVG lithography Micrascan tools.

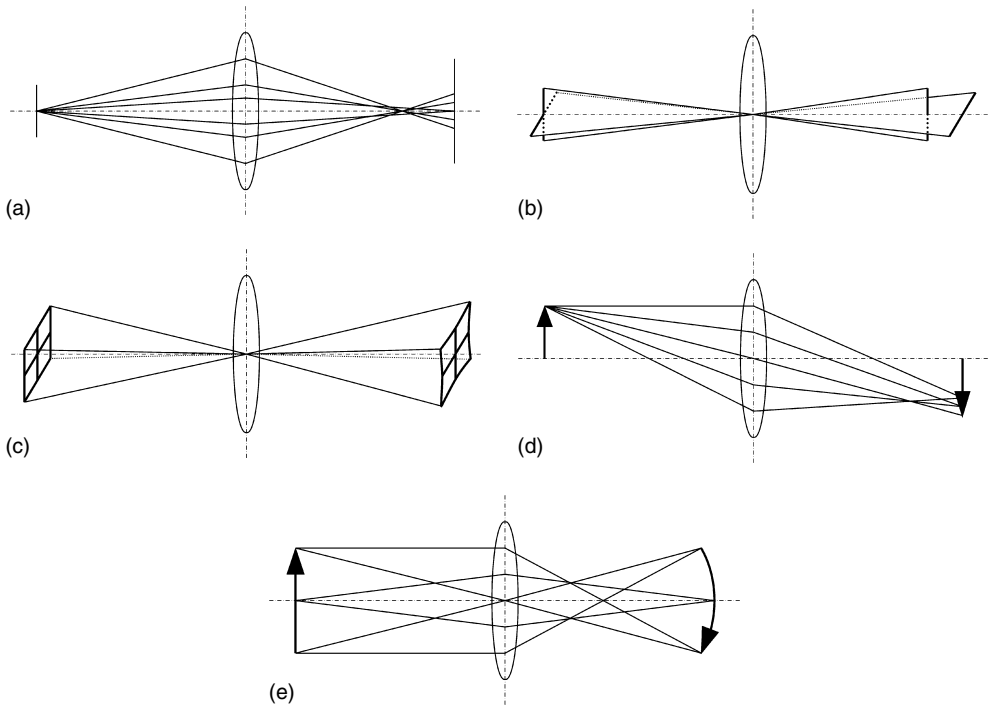


FIGURE 18.7 Primary optical aberrations. (a) Spherical aberration, (b) astigmatism, (c) distortion, (d) coma, and (e) field curvature.

systems, however, the imaging is never perfect. The degradation results from lens aberrations arising from both design and manufacturing of the individual optical components and the complete imaging system [8,9]. It is beyond the scope of this chapter to go into detail on the exact nature and causes of such aberrations, but they are becoming much more important and visible to the practicing lithography engineer in the wafer fab, so a brief discussion follows.

The classic primary aberrations, first analyzed in detail by Seidel in the mid-19th century [10], are sketched in Figure 18.7. These sketches are very simplified and are intended only to give a general exposition of the concept. Extensive details are readily available in many textbooks on optics [11–13].

Spherical aberration describes an imaging defect in which the focal position of the imaging rays traversing through the center of the projection lens aperture is different from the focal position of the rays traversing near the edge of the aperture. The resulting image has lower contrast than the ideal image, and the focal plane position will be seen to vary with changes in the operating NA of the projection lens. One of the practical consequences that can be observed by the lithography engineer is the tilting of the family of CD vs. focus and exposure curves (the so-called Bossung curves [14]) commonly used to analyze and optimize the lithography process.

Astigmatism describes the imaging defect in which the focal plane positions of lines in different orientations are separated. In optics textbook discussions, these lines are typically oriented in radial and tangential fashion with respect to the lens, but in practical lithography the astigmatism is typically described in terms of *x*- and *y*-orientation of the lines. The practical impact is that imaging with an astigmatic projection optical system will often lead to different image linewidths for horizontally and vertically oriented transistor gates and other critical pattern features.

Distortion is an imaging defect in which the magnification of the wafer image compared to the reticle object depends on the distance of the object from the axis of the optical projection system. The significant practical impact is the reduction of image placement (IP) accuracy, leading to overlay degradation. Distortion is one of the major challenges to effective matching of two or more lithography tools for use at different pattern levels on the same wafers.

Coma describes an imaging defect in which the magnification of the image varies with the location of the imaging rays in the aperture of the projection lens. The practical implication of coma for lithography is that resist pattern features become asymmetric and misshapen. A commonly used practical test for coma is the measurement of the difference in linewidth across an array of closely spaced lines. If coma is present the linewidths at the two sides of the array will be different.

The final primary aberration, field curvature, is an imaging defect in which the focal plane position varies with the distance of the object and image from the axis of the optical projection system. The practical impact of field curvature on lithography is reduction in the usable of focus and degradation of linewidth control.

The most common representation of aberrations in use by lithographers is in the form of so-called Zernike polynomials and coefficients. These are described in many optics textbooks, and a very readable description is given in a paper by Brunner [15]. The lowest order terms of interest are astigmatism, coma, and spherical aberration; however, these terms do not have exactly the same definitions as used for the primary Seidel aberrations. The Zernike coefficients are measured with precision interferometry, generally before the projection optical system is installed in the lithography tool. The lowest order coefficients can be measured and analyzed through careful wafer pattern analysis, and these coefficients are widely used in setting up and maintaining exposure tools.

There are a total of 37 Zernike terms generally used, and it has become customary to describe the overall projection system quality as the root mean square (RMS) value of 37 terms. A related quantity, the so-called Strehl ratio, is also used to describe optical systems. For the low aberration levels of interest to lithography, the Strehl ratio can be expressed as a simple function of the Zernike coefficients, a_j,

$$S = \exp\left(-4\pi^2 \sum_{j=2}^{37} a_j^2\right). \tag{18.8}$$

If the aberrations are very small this expression can be simplified to the lowest order terms to become

$$S = \left(1 - \frac{(2\pi \text{ RMS})^2}{2}\right)^2, \tag{18.9}$$

where the wavefront aberration RMS value is expressed in units of λ [16].

As this is written, the leading optical lithography tools available on the market have a Zernike RMS value below 0.005λ, or about a 1 nm wavefront error in the case of 193 nm deep UV lithography. The corresponding Strehl ratio is well over 0.99. It was proposed by Progler [17] only a few years ago that a future "Gold Medal" class lithography projection lens would have a total wavefront error RMS value below 0.025λ. The progress of the lens designers and manufacturers has been very rapid, however, and a state-of-the-art lens today has aberration levels only 20% of the Gold Medal requirement. This rapid advancement is in keeping with the rapidly increasing performance demands for semiconductor manufacturing, as will be discussed later.

One additional aberration can be of considerable importance in advanced lithography. Most lithographic simulation and analysis assume that the wavelength of the exposure light is fixed at a single monochromatic value. The complex high NA lens designs in modern exposure tools are designed to have some tolerance for wavelength variation, but it is desired to reduce that requirement as much as possible so that other type of aberrations can be effectively addressed. The natural bandwidth of excimer lasers is far too broad for use directly, so all lithography lasers contain line-narrowing optics to reduce the bandwidth to 0.5 pm or less. Future generations of exposure tools may require bandwidth of 0.2 pm or less. The role of laser bandwidth and its impact on chromatic aberration is discussed in a publication by Lai [18].

18.4 Exposure Tool System Considerations

18.4.1 Alignment and Overlay

A fundamental requirement for semiconductor lithography is the placement of all pattern edges in precisely the correct location with respect to existing patterns on the wafer. One of the key elements in this placement is the control of the CDs of the pattern features. The other key element is the precise placement of the center of each feature, independent of the size of the feature.

The earliest alignment scheme consisted of a visual microscope inspection of unique features on the wafer and on the adjacent shadow mask in contact with the wafer. The lithography operator adjusted the relative position of the mask and wafer until satisfactory overlay of the unique patterns, known as alignment keys, was achieved. Sub-micron overlay was achievable, but this technique clearly was inadequate for projection optical lithography systems requiring sub-100 nm overlay.

The basic functions of alignment and overlay consist of first detecting a reference mark on the wafer and then positioning the mask and wafer so that the projected image from the mask and projection lens falls on the wafer in the desired location. Alignment and overlay are closely related, but they describe different portions of the total process of placing the images in the correct location. Alignment generally refers to the detection of the special alignment keys on the wafer, and overlay refers to the remaining elements of the image placement, including stage positioning, lens errors, and mask errors.

18.4.1.1 Alignment

Several categories of alignment detection schemes are in use today. One of the earliest and simplest alignment systems consists of a microscope objective attached to the exposure tool lens column and a video camera to receive the image of the wafer alignment key. The microscope objective contains a reference fiducial mark in the form of a crosshair or some other simple shape. Illumination is typically

broadband in a wavelength region longer than the sensitive band of the photoresist. The video output is analyzed by an image analysis computer, and the center of the alignment key is located with respect to the lens column. Ideally, it is then possible to drive the exposure tool stage to the correct location and expose the pattern on the wafer in the correct location.

There are several potential pitfalls in this simple approach. The precision of a simple pattern detection system is generally not good enough to support the sub-10 nm overlay needs of leading edge semiconductor products today. The video imaging is subject to errors from non-uniform illumination and wafer reflectivity variations. The measurement is indirect and does not directly measure the actual positioning of the reticle image through the lens (TTL) onto the wafer. There can be errors in the projection lens that are not detected by the video-based alignment system, and the critical microscope components can move slightly with respect to the optical column. Nonetheless, these off-axis video-based systems have undergone continuous development and they are capable of providing state-of-the-art pattern alignment. The alignment system optical and mechanical components are now manufactured to a precision rivaling that needed in the main projection lens. Simple edge detection techniques have been replaced with sophisticated pattern analysis algorithms [19].

A second alignment detection scheme in wide use depends on diffraction from a special alignment key in the form of a simple grating placed on the wafer. There are numerous implementations of this approach, but in general a low-power laser beam is rapidly scanned over the grating and a set of detectors measures the intensity of one or more orders of the reflected and diffracted intensity peaks as the laser is scanned.

One of the key error sources in any of these alignment mark detection schemes is the uncertainty in the precise location of the alignment reference. In most cases, the reference is indirect, meaning that the actual image positioning is not checked, but rather that a reference outside of the lens is used. This is known as off-axis alignment. In the case of the video image scheme, the location of the microscope objective reference mark is not precisely known and it is not completely stable over long periods. It is necessary to periodically measure the position of the reference through a calibration procedure. This procedure varies for different models of exposure tools, but the result is the determination of a so-called baseline offset for the alignment system. The laser scanning system is subject to the similar sorts of baseline uncertainty, and it is calibrated in a similar fashion.

One approach to reduce or eliminate the baseline error, as well as some of the lens and reticle error components, is to perform the alignment detection directly TTL. Several schemes have been employed for this so-called TTL alignment. In some cases, the alignment light is taken directly from the main exposure illumination source, and in other cases, it is supplied separately from an He–Ne laser or argon ion laser. The detector can be of several forms and can be located in several places on the exposure tool.

The primary benefit of direct TTL alignment is that the baseline error is effectively eliminated. Potentially negative consequences result from additional system complexity, possible obscuration of a portion of the lens aperture, and the use of actinic illumination that can expose the resist on the wafer. In some cases, the TTL alignment is performed with non-actinic red or green light, so resist exposure does not occur during alignment. The projection lens is corrected only for a narrow bandwidth, so the TTL alignment illumination must be carefully engineered. None of the major exposure tool manufacturers today use direct TTL techniques for standard wafer alignment, although TTL methods are used for calibration and baseline adjustment.

Systems using a single alignment wavelength can suffer from significant changes in signal strength due to thin film interference effects in the resist and wafer substrate films. This can lead to alignment degradation or even complete loss of functionality. A solution that has been successfully employed is to use two wavelengths, such as the blue and the green lines of the argon laser. The optical system complexity is increased, but the alignment performance is much more consistent and robust. In some cases, the resist is relatively opaque and/or opaque antireflective coatings are on the wafer, potentially obscuring the alignment keys. It is therefore very important that the alignment scheme selected has high sensitivity to low-contrast images. The video analysis alignment schemes can

generally employ broadband light, and any adverse impact from thin film interference is largely avoided.

18.4.1.2 Overlay

The function of accurately and precisely detecting the alignment keys on the wafer is essential to achieving acceptable pattern overlay, but it is not enough. Several important factors combine with the alignment detection to produce the final pattern placement on the wafer. The most important sources of overlay errors are outlined below.

18.4.1.2.1 Mask Errors

Masks (reticles) for optical lithography are created in a pattern generation process using either an electron beam or a laser beam writer. These tools are highly precise in both CD control and image placement, but residual errors will still exist. Indeed, it is generally accepted that the reticle component to the total lithography error budget is on the order of 40%. Analysis of the sources and characteristics of reticle errors are included in another chapter in this volume, but a few comments are included here. The reticle is a generally robust structure, and once the reticle is manufactured the errors will be quite stable. However, it is important that the reticle is stored and loaded into the exposure tool with care. Small errors in chucking of the reticle onto the reticle stage can lead to apparent distortion on the order of nanometers in the pattern placement on the wafer. Seemingly, minor amounts of contamination can lead to such chucking errors.

18.4.1.2.2 Lens Distortion and Magnification

Distortion is a typical characteristic of any optical projection system. The typical levels of distortion achieved in state-of-the-art exposure tools as this is written are generally below 10 nm and in many cases are approaching 5 nm. It is typical that the distortion is stable, so a given projection lens will not show distortion-induced placement errors unless the pattern is referenced to a previous pattern created by a different lens. This mix-and-match manufacturing approach is very desirable for maximum flexibility and throughput in the wafer fab, so the distortion-induced placement errors must be managed by the lithography engineer. In some of the most critical patterning steps, it is necessary to return to the same exposure tool for all of these critical patterning layers.

Magnification errors are caused by slight deviations in the positions of key elements in the projection lens. All modern projection lenses are telecentric on both the object side and the image side, so slight variations in reticle position or wafer focus should not have an impact on magnification. However, there are small errors in the telecentricity which also lead to magnification errors. All exposure tools have control systems to maintain the magnification (reduction ratio) at the desired value, and this adjustment is typically done as part of the alignment process for each wafer. For this reason, the magnification error is not completely stable in time as was postulated for the distortion error. The typical magnification error is less than 1 part per million (ppm), so it is of no consequence for CD control of sub-micron pattern features. The critical overlay must be maintained over the entire exposure field of 20 mm or more. A magnification error of 1 ppm over a field dimension of 20 mm will yield an image placement error of 20 nm, which is very significant in the most critical applications. Unlike distortion, the magnification error has little impact on mix-and-match operation. The magnification error tends to be random and not characteristic of any one tool in a matched set.

18.4.1.2.3 Stage Errors

In all cases, it is essential that a projection optical exposure tool includes a highly accurate and precise stage to position the wafer in exactly the correct location to receive the projected image. In addition, step-and-scan tools also require a precision stage to move the reticle as needed. These stages are tracked by laser interferometer gauges with measurement resolution of less than 1 nm. Several sources of errors are still possible, including imperfect stage interferometry mirrors used with the laser gauges and

measurement errors induced by turbulence or temperature non-uniformity of the air in the region of the stages. Modern exposure tools require very careful design of the airflow and thermal loading around the stages, with the main reason being the need for accurate laser interferometry. The typical errors due to stage position measurement are a few nanometers.

The measurement of the stage position is critical, but even with a perfect measurement it is not possible to position the stage exactly as needed. Stages are large complex mechanical systems, and the movement dynamics are very important. Key factors are friction, stiction, hysteresis, and vibrational modes. Frictional forces have been almost completely eliminated by the widespread use of air-bearing stages, but other sources of mechanical drag are still possible. In the case of scanning exposure tools, an additional complication arises from the need to carefully synchronize the movement of the reticle and wafer stages. An important consideration for any exposure system is maintaining high throughput. This in turn drives the need for high stage speeds and rapid accelerations, and it strongly constrains the allowable time for settling and vibration damping. It is sometimes possible to make a trade-off in the wafer fab between throughput and stage precision, depending on the needs of the wafer product. Numerous papers on precision stages have been published, and the interested reader is directed to the references for more information [20,21].

18.4.1.2.4 Wafer Distortion due to Thermal Processing

A classic error source in lithography is the overlay error associated with distortion of wafers. Extreme or improper heating of silicon wafers leads to slip in the silicon and the subsequent movement of various features on the wafer surface. The difficulty in correctly overlaying the existing pattern with a new imaged pattern can be severe. The author has experienced cases in which the wafer distortion was found to approach 1 μm following experimental thermal processing in a development laboratory. Most thermal processes in the wafer fab are very well controlled to maintain the electrical characteristics of the wafer, so it is not common to observe related overlay problems. However, this distortion should be considered when an otherwise mysterious overlay problem occurs.

18.4.1.2.5 Wafer Distortion due to the Addition and Removal of Highly Stressed Films

Most thin film layers used on semiconductor layers are deposited with some residual stress, either compressive or tensile. In some cases, this stress can be quite large, leading to measurable changes in the wafer dimensions when the wafer is chucked onto the exposure tool wafer stage. Thick oxidation layers, such as those formerly used for the field isolation in metal-oxide semiconductor devices, can drive wafer dimensional changes of several parts per million. In most cases, the stress from a uniform film will result in a linear dimensional change that is readily and automatically compensated by the magnification control of the exposure tool.

A different problem results when the highly stressed film is partially removed from the wafer, as during the definition of interconnect structures on the wafer. If the resulting interconnect pattern is not uniform across the wafer, or at least across the chip, then the stress-induced distortion will not be linear. In this case, it may not be possible to correct for this distortion through any adjustments of the exposure tool. A significant amount of research work has been performed by Engelstad and co-workers [22,23] on this topic. The driver for this work was primarily next generation lithography approaches, which would use thin membrane masks, but many of the same stress-related distortion issues also apply to silicon wafers and optical lithography.

18.4.1.2.6 Apparent Displacement of the Wafer Alignment Keys by Asymmetrical Resist Coating

Alignment targets on the wafer typically consist of raised or depressed features in the underlying pattern layers. The vertical size of these targets is therefore the same as the thickness of the underlying substrate layer, often a few hundred nanometers up to a micron or more. The photoresist is invariably applied in a spin coating process. The coating of the resist is not completely conformal to the substrate nor completely

planar, and the resist tends to coat the alignment targets asymmetrically. This asymmetry shifts the apparent position of the alignment target as represented by it surface topography. The asymmetry in the resist coating depends on the alignment target wafer location and its orientation, and therefore the potential error varies across the wafer. In some cases, the location with respect to the chip may also be important, since the surrounding pattern density will affect the coating properties over the alignment target. Depending on the exact nature of the alignment detection system in use, the detected target position may be shifted due to this asymmetry [24]. Modern alignment software algorithms are designed to capture and compensate for this asymmetry, but some error may remain. Planarization of the wafers during chemical mechanical polishing (CMP) processing greatly reduces the resist coating asymmetry for some critical pattern layers. However, as describe next, CMP can lead to certain errors as well.

18.4.1.2.7 Displacement of Wafer Alignment Keys by CMP of Metal Layers

A displacement in the alignment target positions can be found following CMP of metal layers, which is now in widespread use. In this case, however, the alignment target itself, and not just its apparent location, is physically altered by the mechanical action of the polisher. In the typical case, the alignment targets are rounded more on one side than the other, leading to asymmetrical marks and an effective displacement of the centroid. In more extreme cases, the target can actually be moved by a smearing action from the CMP tool. The nature of the mark displacement is not uniform across the chip or even the wafer. Instead, the CMP action typically causes a swirl pattern on the wafer that is not centered at the center of the wafer. Chemical mechanical polishing process optimization eliminates most of the mark shift problem, but it is still a potential source of error.

18.4.1.2.8 Wafer Chucking Errors

All optical lithography exposure tools include a provision for flattening and clamping the wafer onto the wafer chuck located on the wafer stage. If this chucking is not performed correctly, typically due to contamination on the wafer or the chuck, then the wafer will not be flattened to its standard configuration. This leads to the well-known focus "hot spot" problem that causes resolution failures and CD errors. In addition, the effective locations of wafer features with respect to the optical column are slightly changed due to local tilt of the wafer as it rests on the contamination.

18.4.1.2.9 Overlay Metrology Errors

The difficulty in achieving pattern overlay of a few nanometers is also shared by the tools and techniques used to measure the overlay. The most common overlay measurement artifact in use today is the so-called box-in-box structure. A large (10 μm) box or box-like pattern is printed on the wafer during the imaging of the first pattern of the overlay pair, and a second box, of slightly different size, is printed during the exposure of the second pattern. There are several commercially available metrology tools that can measure the relative positions of the two boxes to within a few nanometers. The potential metrology errors arise from several sources. First, there are various offsets and shift in the metrology itself. These are generally below about 2 nm so they are relatively minor. More importantly, the structure and dimensions of the boxes can be a significant factor in measurement errors. The pattern features comprising the boxes are generally much larger than the pattern features of interest in the chip area of the wafer. Aberrations in both the exposure tool and the metrology tool can lead to offsets in the measured box positions from the critical wafer feature positions. Finally, the overlay metrology structures are generally placed in the scribe line areas between the product chips. Any non-uniform overlay error that varies across the chip may not be determined correctly by structures in the scribe lines.

18.4.1.3 Overlay Analysis

One of the primary reasons for the introduction of wafer steppers as a replacement for full-wafer projection printers was to improve the overlay of pattern levels. The biggest sources of errors in full-wafer

tools were mask errors and wafer distortion errors due to processing. Steppers minimize the mask errors by allowing an increase in mask feature size, thereby reducing the impact of placement errors of a given size. Steppers also allow local positioning in a small area on the wafer, thereby reducing the impact of process-induced wafer distortion.

The first widely used model for describing the overlay errors associated with stepper lithography was published by Perloff [25]. Since most circuits are designed in an orthogonal fashion using the so-called "Manhattan geometry," the error components and resultant sums are typically calculated separately for x and y components. The lowest order terms are translation, rotation, and magnification. Simple expressions for overlay errors are:

$$\Delta x = T_x - \theta \times y + M_x \times x + \text{residual}, \tag{18.10}$$

$$\Delta y = T_y + \theta \times x + M_y \times y + \text{residual}, \tag{18.11}$$

where the residual components contain several additional sources of error, including lens distortion and trapezoidal magnification errors.

In the case of step-and-repeat tools, the error components result from both the full-wafer errors, generally called interfield errors or grid errors, and the individual chip errors, generally called intrafield errors. For example, the entire grid of chips may be scaled, rotated, and shifted with respect to the underlying wafer pattern. In addition, each chip may also be subject to scaling and rotation. Full-wafer errors are generally caused by wafer rotation on the stepper stage and stepping size errors. Individual chip errors are caused by reticle rotation and lens magnification errors.

Scanning exposure tools add extra degrees of freedom, which permit additional types of errors and allow additional means of correction. In particular, the magnification for the x and y components can be independently adjusted. There are also two effective intrafield rotation components, one due to reticle rotation and the other due to non-orthogonality of the scan axis to the underlying wafer grid. Figure 18.8 shows several of the more important overlay error components.

In general, the overlay error components can be separated into systematic and random portions. A key question that often arises is how to predict the total overlay behavior, given a set of both random and systematic components. The classic technique of combining random (Gaussian) distributions through the root sum square (RSS) analysis is often used due to simplicity, but in many cases the resulting error estimate will be too low. On the other hand, a simple addition of the error terms will almost certainly be too pessimistic. The most widely accepted approach today is to combine the random errors in an RSS fashion and then add the systematic components in a simple algebraic manner.

The selection of the target pattern layers for the alignment sequence can be important, especially in the earlier portions of the wafer fabrications process. For example, the contact pattern generally needs the best possible alignment to both the underlying active area and polysilicon gate patterns. The alignment to either of the underlying patterns can be direct, but the alignment to the other then becomes indirect and subject to larger errors. Recent advances in exposure tool capability now allow the simultaneous alignment to two different underlying layers. For example, a contact layer could be aligned to the active area in the x-direction and the polysilicon gate layer in the y-direction.

The choice of alignment strategy is made by the lithography engineers and the circuit layout engineers to achieve the best performance and yield for a given product. An analysis of the issues involved in this alignment tree decision has been made by Kopp [26]. In the consideration of alignment sequences, it is common to consider the alignment errors to be random and uncorrelated, so the result of an indirect alignment such as described here is an error that is larger than the basic alignment error term by a factor of the square root of 2. In practice, the errors are seldom random, but this is a good starting point for considering alignment sequences.

This error combining approach, while appearing straightforward, is subject to some potential confusion in the classification of the errors. For example, the alignment of the reticle in the exposure tool has a largely random nature, both in offset and rotation. However, once the reticle is loaded, every

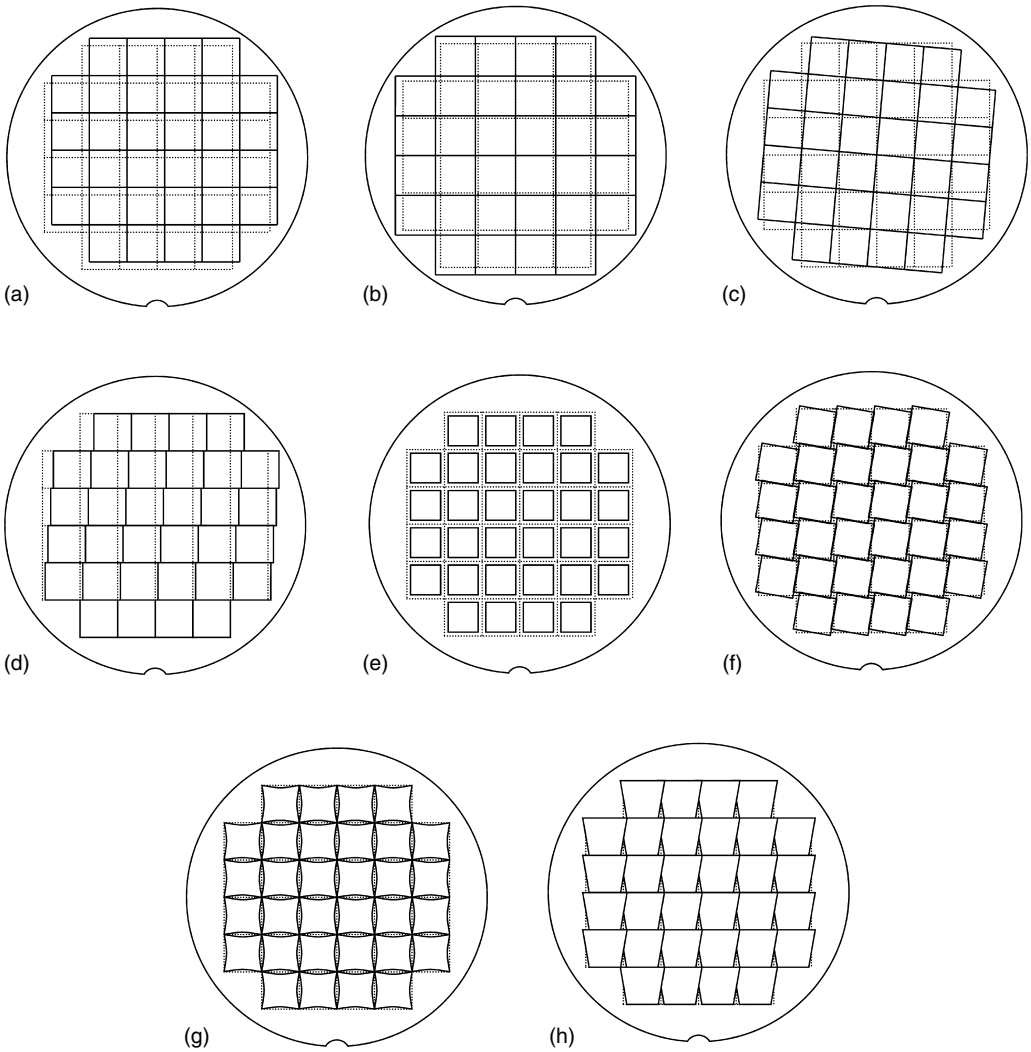


FIGURE 18.8 Overlay errors. The dotted lines represent the ideal grid, and the solid lines show the actual chip positions. (a) Global offset, (b) global scaling, (c) global rotation, (d) orthogonality, (e) field magnification, (f) field rotation, (g) distortion, and (h) trapezoid.

wafer and every chip will experience the reticle loading error as a systematic error. This is a well-known behavior in the realm of statistical analysis. The statistical treatment of errors depends on the population being considered. Over a long period, the impact of reticle loading error on wafer overlay could be considered random, but on a wafer-by-wafer basis the error is systematic.

The difficulty in unambiguously treating error components has led to the development of refinements in characterizing the total overlay. The most common description for total overlay is the mean offset plus the three-sigma value of the (assumed) random error from all other sources. The underlying assumption is that the mean offset can be driven to small values through careful measurement and adjustment of various baseline components in the alignment system and exposure tool stage.

Another common description of the total overlay characteristic is the so-called "good fields rule." In this case, a maximum allowable overlay is specified, and the percentage of stepper fields that completely meet this specification is declared good. This description format is particularly useful when the error components are not normally distributed or when they are not well understood.

A longstanding "rule of thumb" is that the total product overlay error should be no greater than one third of the minimum pattern half-pitch. The total product overlay includes all components in the wafer process, both lithographic and non-lithographic. The overlay error component allowed for lithography is somewhat smaller, at about 20% of the minimum half-pitch. The International Technology Roadmap for Semiconductors (ITRS) roadmap described later uses this rule, as shown in Table 18.3. The overlay error specification directly drives many circuit layout rules, including contact to gate spacing, gate extension beyond the active area, active region edge to complementary metal-oxide-silicon (CMOS) well boundaries, metal to contact overlaps, and so on.

In some cases, additional process complexity can provide a degree of "self-alignment" that can ease the overlay tolerance requirement. An example is the use of a silicon nitride film on sidewall of a polysilicon gate feature. Subsequent contact patterning and etching will not penetrate the nitride layer, even if the desired overlay tolerance is not achieved in the lithographic exposure process. The cost of this self-alignment is additional process steps, including pattern layers, and in some cases the extra film can cause degradation of circuit performance through added electrical capacitance.

Larger overlay specifications provide easier wafer fab lithography operations, but the chip sizes quickly become larger as the overlay tolerance is increased. Smaller overlay tolerances are beneficial for chip size reduction, but the yield of good chips decreases rapidly as the exposure system capabilities are exceeded.

An analysis by Arnold and Greeneich [27] shows that the good chips per wafer as a function of overlay tolerance starts at zero for zero overlay tolerance, increases to a maximum, and then decreases as the tolerance grows larger. The falloff at large tolerances is due to the reduction in the number of chips that can fit on the wafer. Even though the yield may continue to improve the resulting number of good chips decreases.

18.4.2 Throughput

Photolithography is generally the most used process in the wafer fab, typically accounting for 30%–35% of the total process cost. Therefore, most fabs are designed so that the photolithography area will be fully loaded at all times. This in turn drives very high importance to the productivity of the photolithography tools. The wafer throughput is the most prominent factor in advanced cost of ownership models, even more important than the price of the tools.

The throughput capability of an exposure tool results from the combination of numerous factors. The simplest expression of the maximum run rate, sometimes called the "sprint rate" is:

$$T\ (\text{wph}) = \frac{3600\ (\text{s/h})}{\text{wafer process time (s)}}, \tag{18.12}$$

with the wafer process time described by:

$$t_{\text{wafer}} = t_{\text{woh}} + N_{\text{f}}(t_{\text{exp}} + t_{\text{foh}}), \tag{18.13}$$

where t_{woh} is the overhead time per wafer, t_{foh} is the overhead time per exposure field, t_{exp} is the actual exposure time during which the actinic radiation is exposing the resist, and N_{f} is the number of exposure fields per wafer.

The overhead and exposure times contain a number of subcomponents detailing wafer loading, alignment, unloading, stepping, overscan, shutter delay, and stage settling times. In turn, these components can be further broken down into basic machine operations.

The achieved throughput in a manufacturing environment depends not only on the sprint rate capability of the exposure tool but also on the throughput capability of clustered resist processing tracks,

lot setup times, reticle change times, scheduled and unscheduled maintenance, and the availability of wafers to be processed. It is customary to include in the total throughput calculation the time related to lot and reticle changing, as well as any time delays in the movement of wafers between the exposure tool and the resist track. Delays due to maintenance, setups, and wafer availability are generally treated separately as a part of an overall factory capacity model.

The net throughput of the exposure system can be described in the following hierarchy.

18.4.2.1 Exposure Field

The exposure time per field, t_{exp}, is defined for steppers and scanners as:

$$t_{exp}\text{stepper (s)} = \frac{\text{exposure power (mW/cm}^2\text{)}}{\text{resist sensitivity (mJ/cm}^2\text{)}} \tag{18.14}$$

$$t_{exp}\text{scanner (s)} = \frac{\text{field length (cm)} + \text{slit width (cm)}}{\text{scan rate (cm/s)}} \tag{18.15}$$

The scan rate is limited either by the tool manufacturer's mechanical specification or by the resist requirement, whichever is smaller. The mechanical specification provided by the leading exposure tool suppliers at this time is 50 cm/s or greater. The resist driven scan rate limitation is created by the need to fully expose the resist during the time in which the illumination field passes over any given site on the wafer. It can be calculated as:

$$\text{Scan rate (cm/s)} = \frac{\text{slit width (cm)} \times \text{power density (mW/cm}^2\text{)}}{\text{resist sensitivity (mJ/cm}^2\text{)}} \tag{18.16}$$

It can be seen that the optimum slit width includes a trade-off with respect to throughput. A smaller slit reduces the need to overscan the image field on the wafer, and a larger slit eases the scan rate restriction due to resist sensitivity. The slit widths chosen by the tool suppliers at this time range from 5 to 8 mm.

For a typical 193 nm chemically amplified resist, with an exposure sensitivity in the 20–30 mJ/cm^2 range, and an exposure power density of 2000 mW/cm^2 at the wafer, the resist-limited scan rate is 30–80 cm/s, which spans the mechanical scanning rate specification.

The factor t_{foh} includes stepping time, settling time, scan acceleration time, die-by-die alignment time (if used), and shutter delay times. These overhead terms consume typically on the order of 100–200 ms. An example of the relative throughput as a function of field size and resist sensitivity is shown in Figure 18.9.

18.4.2.2 Full Wafer

The full-wafer exposure process time, t_{woh}, includes the wafer loading and unloading time and the wafer alignment time. The load and unload times are determined primarily by the system design, and are generally a few seconds or less. The wafer alignment time is a function of several choices in alignment strategy. The simplest alignment scheme is to do no alignment at all. This is a common case for the first-level pattern, but is not adequate for any other pattern layers. The next approach is to perform an alignment to two or three marks widely spaced on the wafer. The most common approach today is to perform alignment at 8–10 positions on the wafer, including two or three places in each of several fields. This allows detailed calculations of a number of overlay error terms and the appropriate corrections. A unique approach to reducing the wafer overhead time is followed by ASML, one of the major tool suppliers. In these tools, two interchangeable wafer stages are used alternately. During the actual exposure process on a wafer held on one wafer stage, the second stage supports wafer exchange and alignment. When the first wafer exposure is completed the wafer stages are quickly exchanged and the second wafer can begin the exposure process almost immediately [28].

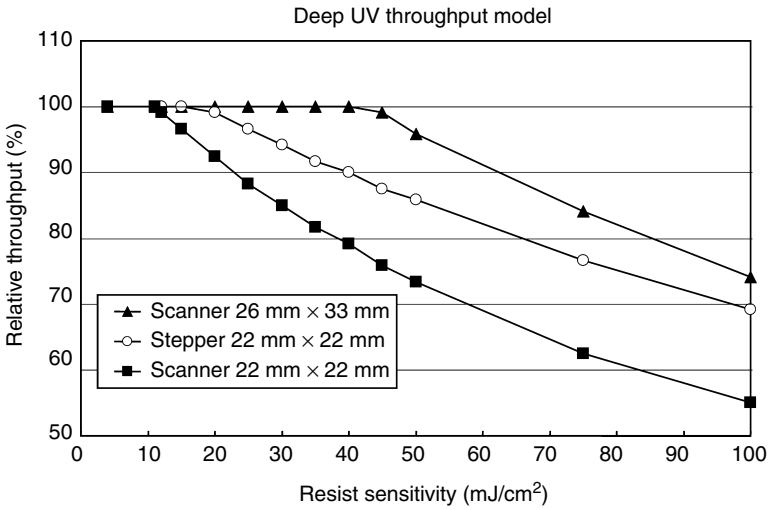


FIGURE 18.9 Throughput model for steppers and scanners with varying field size and resist sensitivity.

18.4.2.3 Production Lot

Additional time is typically required when each lot of wafers is patterned. In the simplest case of a high-volume product, there may be no additional delay if the reticle and process recipe remain the same between lots. In the case of lower volume products, there is additional time required to change the reticle and initiate a new process program. In some cases, there is also a delay due to the required changeover of the resist process on the track.

18.5 Resolution Enhancement Techniques

The current level of optical exposure tool manufacturing has reached very high standards of perfection. There is virtually no opportunity for further increases in NA for conventional projection lenses, and the aberrations are shrinking. Three possible approaches to achieving greater resolution are to reduce the wavelength, to perform the exposure in a liquid environment, and to improve the image formation capabilities of the existing lithography tools and technology, as represented by a reduction in the Rayleigh k_1 factor.

The progression to ever smaller wavelengths has already been discussed. At this writing, the leading edge capability in the industry employs an exposure wavelength of 193 nm. A significant amount of research was undertaken in the late 1990s and early 2000s to develop the next optical wavelength generation, proposed to be 157 nm. This wavelength is generated by an excimer laser based on F_2. The excimer laser has proved to be a suitable light source for the 248 and 193 nm exposure tools now in widespread use, and the extension to 157 nm (F_2) was anticipated to work in a manufacturing-worthy fashion.

However, the challenges of developing the required materials and systems for 157 nm lithography proved too severe to be economically viable. Projection imaging techniques used for optical lithography almost always require at least some transmitting optical elements in addition to some possible reflecting elements. There are very few materials that transmit even at 157 nm, and almost none are available for shorter wavelengths. The primary choices at 157 nm are CaF_2 and BaF_2, with CaF_2 being considerably more mature as an optical material. Unfortunately, CaF_2 has a large intrinsic birefringence property at 193 nm, which significantly complicates the design and manufacture of high-quality optical lenses [29].

In addition, the manufacture of suitable quality CaF_2 was found to be difficult and expensive. There were equally severe challenges for resist materials and for reticle materials. At this time, it does not appear that wavelengths shorter than 193 nm will be used for optical lithography. Extreme UV lithography (EUVL) using a radically smaller wavelength, 13.6 nm, is actively under development at this time. It is photon-based, and therefore also known as "optical lithography." However, further discussion of EUVL is outside the scope of this chapter.

Resolution enhancement by performing exposure in an environment other than air, specifically in water, is a technique that is in an active development and implementation phase at this time. Discussion of immersion lithography is in Section 18.7.

The third alternative for resolution improvement in high-volume manufacturing is to apply one or more of the so-called resolution enhancement techniques (RETs). Imaging performance can be enhanced in several ways. Three overall types of improvements include wavefront engineering, which attempts to custom tailor the aerial image to provide increased resolution [30], mask engineering, which optimizes the exact shape of patterns on the mask to provide the desired patterns on the wafer, and resist process engineering, which optimizes the ability of the resist system to receive the aerial image and produce the desired features on the wafer. Table 18.1 shows the main categories of RETs currently being researched and used.

The various RET approaches have been heavily researched in the past few years, and numerous papers have been presented in the leading lithography conferences around the world. In addition, all of these techniques, with the exception of pupil filtering [31], are now being introduced into high-volume wafer fab manufacturing. The RETs interact in many ways, and it is important to consider those interactions when designing and refining the lithography process. Smith has discussed this requirement in some detail in a recent publication [32]. In the present chapter, the RET methods will be treated separately. The first three topics in Table 18.1 will be described further below.

18.5.1 Phase Shift Masks

Many concepts in optics used for photolithography go back over 100 years. Phase shift mask technology for optical lithography builds on the principle of interference between light waves. In conventional lithography, the light from adjacent openings in the mask overlaps in the dark region between the mask

TABLE 18.1 Resolution Enhancement Techniques (RETs)

RET	Type	Advantages	Disadvantages
Phase shift masks	Wavefront engineering	Improves DoF and exposure latitude	High mask cost, inspection and repair difficult
Modified illumination	Wavefront engineering	Improved DoF for dense line/space features	Less improvement for holes or isolated lines affected by lens aberrations
Optical proximity correction (OPC)	Mask engineering	Improved critical dimension (CD) control for various size patterns	Additional design data processing masks more complex and expensive
Wafer control—antireflective layers	Resist engineering	Improved CD control reduces notching	Increased cost and process complexity may complicate etch
Pupil filtering	Wavefront engineering	Improved CD control and exposure latitude	Pattern-specific capability must be designed in by lens manufacturer
Multilayer and surface imaging resists	Resist engineering	Improved CD control, improved resolution, includes antireflective functionality	Increased process complexity and cost, generally requires plasma etch capability as part of photolithography

openings and thereby lowers the contrast between the light and dark regions. If the phase of the electromagnetic waves comprising the light in the adjacent mask openings is different by 180° then there is destructive interference between the light from the adjacent openings, and the resulting image is dark between the open features. This concept has been used for a long time in various non-lithographic optical applications, such as microscopy, and it was first described by for lithography use in terms of x-ray lithography.

The first use of phase shifting masks (PSMs) for optical lithography was demonstrated by Levenson and co-workers at IBM in the early 1980s [33]. Independent development of phase shifting was also underway in the same timeframe by Shibuya [34] and Smith [35,36]. The general concept of the so-called alternating phase shift mask is shown in Figure 18.10. The phase shifting effect in the mask originally was accomplished by adding a thin layer of transparent material, typically a spin-on glass, in the correct thickness and then patterning this layer to create the phase shift where needed. The correct thickness of the shifter is simply the physical thickness that provides an optical path length exactly one-half wavelength longer than the optical path length in the same thickness of air. The index of refraction of standard air is 1.0003, so the wavelength of light in air is very close to the vacuum wavelength, and this factor is generally ignored in the calculations for phase shifting layers. The required phase shifter thickness is given by:

$$t = \frac{\lambda}{2(n-1)}, \tag{18.17}$$

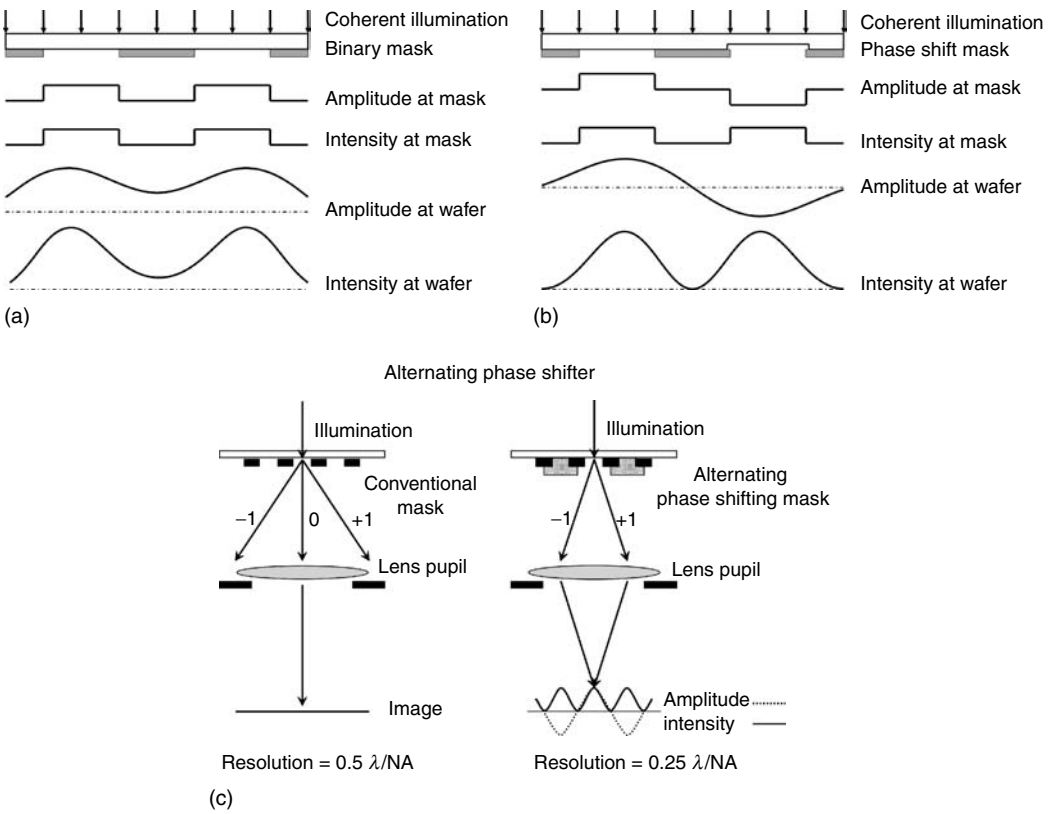


FIGURE 18.10 Alternating phase shift masks. (a) Superposition of aerial image amplitudes for coherent illumination of a binary mask, (b) superposition of aerial image amplitudes for coherent illumination of an alternating phase shift mask, and (c) sketch of resolution improvement mechanism.

where n is the index of refraction of the shifter material. For typical conditions with $n = 1.5$, the phase shifter thickness is the same size as the exposure wavelength.

The precision and quality of spin-on films is not adequate to provide phase shifting to within 1 or 2%, so the practice of making phase shift masks has evolved to etching the required phase-shifted pattern features directly into the mask substrate. The mask substrate is almost always high-quality fused silica, and the material properties are very well controlled. The control of the etching follows from common wafer fab practice and is also well controlled.

In practice, however, the structure of phase shift masks must be more complex. The simple structure described above leads to unequal light intensity in the two phase regions, 0 and 180°, and the inequality varies with feature size and with focus. A detailed analysis through rigorous electromagnetic wave modeling of the optical behavior in such a mask shows that it is necessary to adjust the etch sidewall profiles and/or bias the size of the pattern features to achieve the full benefit of the phase shift effect [37].

It is outside the scope of this chapter to discuss details of mask fabrication, but it is clear that the manufacturing of these masks is considerably more difficult than the manufacturing of conventional chrome-on-glass masks. An addition pattern write step is needed, along with the subsequent etch, and the inspection and repair of these masks is much more complex.

A major concern in the use of alternating phase shift mask technology is the limitation on application to an arbitrary circuit pattern layout. As shown in Figure 18.11, this is a topological problem and not simply a lithography technology challenge. In many cases, the pattern layout requires that opposite phases butt together, creating an unwanted dark line in the image on the wafer. In other cases, there is an ambiguity in selecting the phase to be applied to specific openings. If two adjacent openings have the same phase then the resolution enhancement benefits will be absent in this region. Numerous approaches to modify the pattern layout to allow alternating phase shifters have been developed, but no completely general solution has been reported that does not also increase the chip area.

One approach, first proposed by Levenson, is to reverse the tone of the mask to use negative tone resist rather than the more common positive tone resist. Typical circuit patterns consist of separable line features that can be phase shifted if they are designed as clear openings on the mask, and the exposure is made into negative tone resist. The background pattern area is usually connected across the chip and cannot be phase shifted in a useful manner. The use of negative tone resist often removes the direct phase conflict abutment errors, but it does not solve the ambiguity problem. The only known general solution to the ambiguity problem is to increase the distance between the conflicting pattern geometries so that phase shifting is no longer needed to provide image resolution. Software programs have been developed, which assign phases and adjust pattern positions where phase ambiguities exist [38].

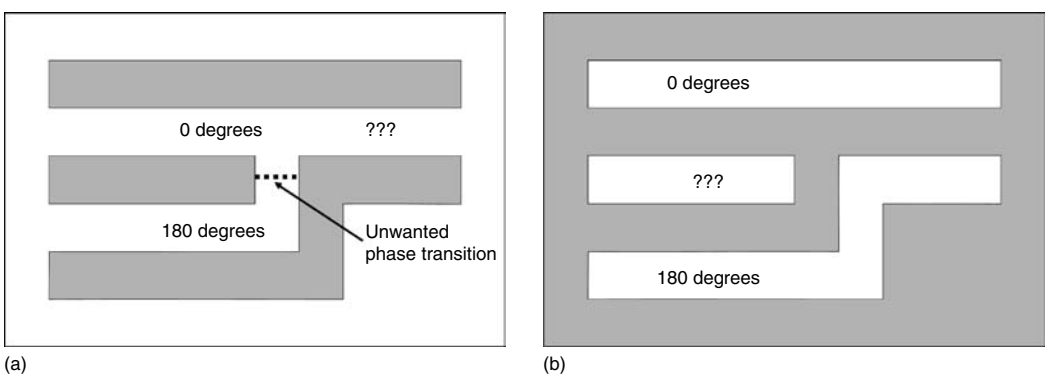


FIGURE 18.11 Topology dilemma for alternating phase shift masks. (a) Bright field case and (b) dark field case.

A different approach to avoiding phase conflict problems is to split the complete pattern into two separate mask layers. The most critical features, such as polysilicon gate patterns, are placed on an alternating phase shift mask, whereas the less critical features, such as the polysilicon interconnect patterns, are placed on an ordinary binary mask [39,40]. In this manner, the phase conflicts can generally be avoided, and the most critical pattern features receive the benefit of the high imaging performance provided by phase shifting. The penalty is that two masks must be fabricated and exposed, increasing the cost and complexity of the overall patterning operation.

A significantly different implementation of the phase shift mask concept was developed by Terazawa and co-workers at Hitachi [41]. This approach uses a nearly conventional mask structure with the opaque chrome pattern layer replaced by a composite phase shift layer. The composite pattern layer provides the 180° phase shift and, in addition, allows only a few percent of the light to pass through the phase-shifted regions. This so-called half-tone or attenuated mask is much easier to manufacture than the alternating phase shift mask, and it has seen considerable application for contact and via patterning as well as other pattern layers. The current favored implementation uses a single layer, typically a non-stoichiometric MoSi film, to simultaneously provide the required phase shift and attenuation. This mask form is often called the embedded attenuated phase shift mask (EAPSM).

The attenuated PSM functions with the same basic interference principles as the alternating PSM, but the details are quite different. Sketches of the attenuated PSM enhancement mechanism are shown in Figure 18.12. The intensity of the phase-shifted light is only a few percent of the intensity of the unshifted light, there is no zero-intensity zone in the attenuated PSM image as there is in the alternating case. The optimum amount of transmission is a trade-off between maximizing the phase shift effect and minimizing the unwanted artifacts from light leakage through the attenuating layer in areas between the desired pattern features. A particularly important artifact can occur at the location of the first intensity peak in the Airy function described previously. This light at this peak has a 180° phase shift from the central Airy peak, so the light from the first peak adds to the phase-shifted light leaking through the attenuated mask. This addition of light from both the Airy diffraction and the EAPSM transmission can create unwanted artifacts called sidelobes [42]. The sidelobes will expose the resist sufficiently that unwanted defects in the resist will be formed. An example of a severe case is shown in Figure 18.13. The solution to this problem requires both careful selection of the transmission factor and in some cases pattern layout adjustment to avoid the direct overlap of multiple sidelobes from multiple pattern features.

Several other implementations of the phase shift mask concept have been developed over the past 20 years. A brief description of the most noteworthy forms is given in Table 18.2.

18.5.2 Modified Illumination

A fundamental requirement for projection imaging is that at least the lowest order non-trivial components of the spatial Fourier transform of the reticle image must be captured by the pupil of the projection lens. This can be viewed in an instructive fashion as shown in Figure 18.14. For direct imaging of a reticle illuminated with spatially coherent light, i.e., the illumination beam consists of parallel rays perpendicular to the reticle, the first-order diffraction peaks from the reticle must be captured within the pupil of the lens. This leads to a simple resolution limit of $0.5\lambda/\mathrm{NA}$, or in equivalent terms a k_1 limit of 0.5.

In the case of partially coherent illumination, the illumination beam has an angular spread related to the partial coherence factor, sigma [43,44]. The maximum useful sigma is 1.0, at which point the illumination beam just fills the lens pupil without any diffraction or scattering by the reticle. As can be seen from Figure 18.15, this illumination condition can theoretically double the resolution capability of simple grating type mask patterns, since portions of the first-order diffraction peaks from the reticle can be captured even when the nominal diffraction angles are twice as large as the limiting coherent case. In practice, this resolution doubling is not achieved since the amount of light transmitted is small, the depth of focus is small, and the typical mask pattern has much more variety than simple gratings.

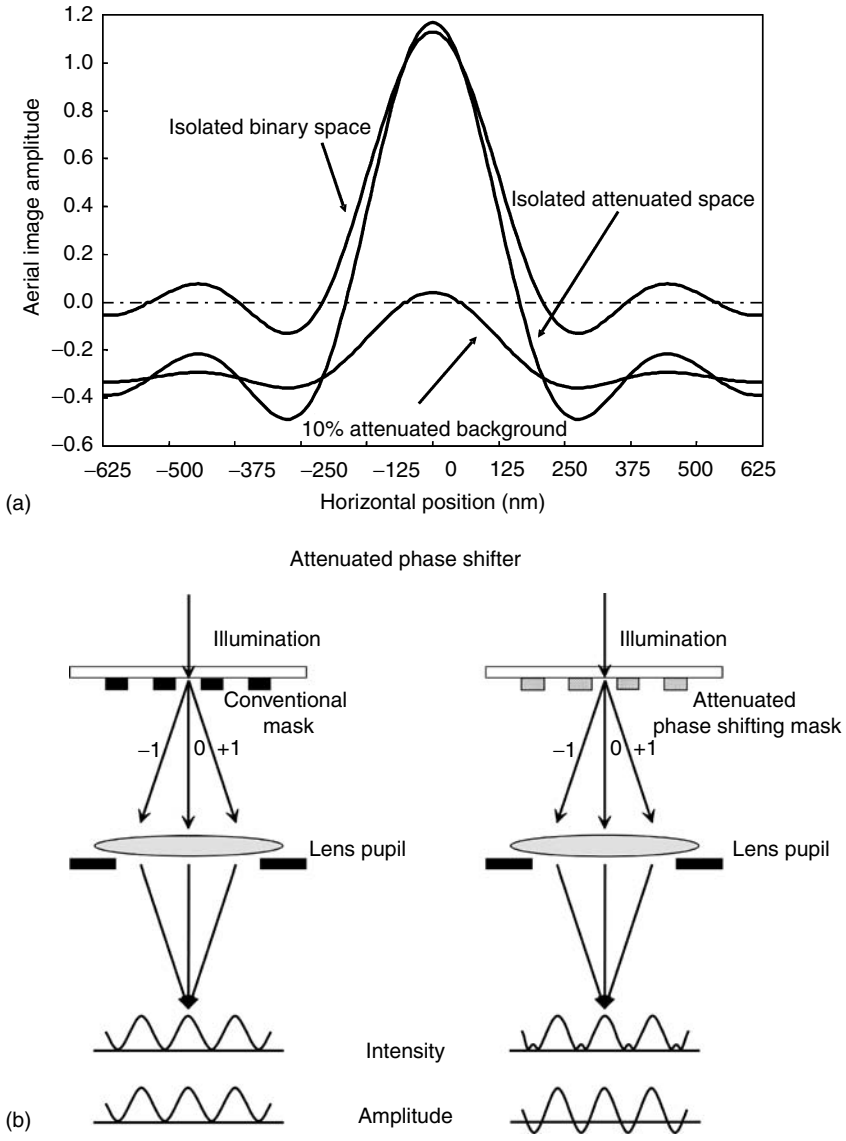


FIGURE 18.12 Attenuated phase shift masks. Transmission through dark phase-shifted areas of mask is 10%. (a) Superposition of aerial image amplitudes for coherent illumination and (b) sketch of resolution improvement mechanism.

The resolution limit in terms of the Rayleigh k_1 factor can be modified [45,46] to include the effect of partial coherence as expressed by

$$k_1(\text{limit}) = \frac{1}{2(\sigma + 1)}. \tag{18.18}$$

Diffraction-limited resolution ranges from $k_1 = 0.25$ for incoherent illumination to $k_1 = 0.50$ for coherent illumination.

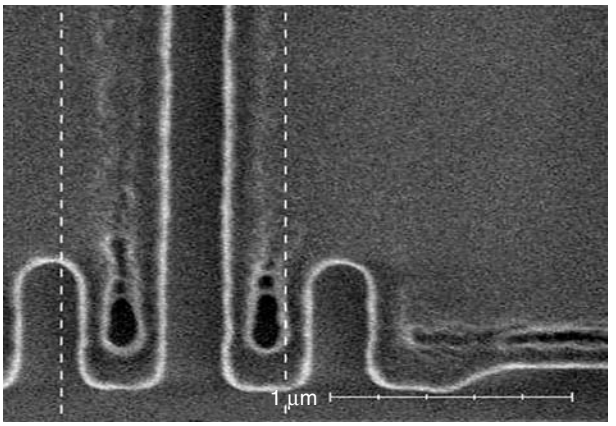


FIGURE 18.13 Unwanted exposures resulting from sidelobe light leakage through attenuated phase shift mask.

The apparent resolution improvement obtained from large sigma values must be carefully considered with respect to other important factors. The edge profile of the aerial image is sharper with coherent illumination than with incoherent illumination. Thus, the resolution may improve with high sigma, but the CD control can be worse in a practical application. Clearly, it is important to perform a thorough analysis of the complete imaging performance to determine the best value for sigma.

The configuration of the illumination profile can be extended to additional structures such as annular rings, dipoles, quadrupoles [47–49], and various combinations. Many of these configurations operate in a manner that the illumination is primarily or completely off-axis. The use of off-axis illumination improves resolution by allowing at least one of the first-order peaks, along with the zero order, to pass TTL and form the image. A simple case of off-axis illumination is sketched in Figure 18.16. Again, in the extreme case the benefit provided by off-axis illumination would be a doubling of the resolution capability of the imaging system compared to a fully coherent on-axis illumination scheme.

TABLE 18.2 Phase Shift Mask Formats

Type of Phase Shifting Mask (PSM)	Comments
Alternating	The original Levenson type described in the main text
Attenuated or half-tone	The attenuated mask is the most widely used PSM today. Described in the main text
Chromeless	This is a special case in which lines are defined by the edge transitions between 0 and 180° phase regions. Creates unwanted artifact features and must be used with a second mask to trim out the artifacts
Rim shifter	The enhancement is performed by a narrow rim of phase-shifted area surrounding the main pattern feature. The advantages are simpler mask manufacturing (only one pattern) and the absence of layout topology limits
Outrigger	Each pattern feature is "self-contained" with the phase shifting performed by a pair of small shifted openings adjacent to the main feature. Can also be used for contact patterns, with four surrounding shifted openings. In some cases, the outriggers can be shared between features
Multiphase	An adaptation of the alternating PSM in which unwanted phase transitions are smoothed out through the use of a two- or three-step transition. For example, the transition between 0 and 180° contains small regions of 60 and 120° phases. This eases the topology problem and avoids the need for trim mask

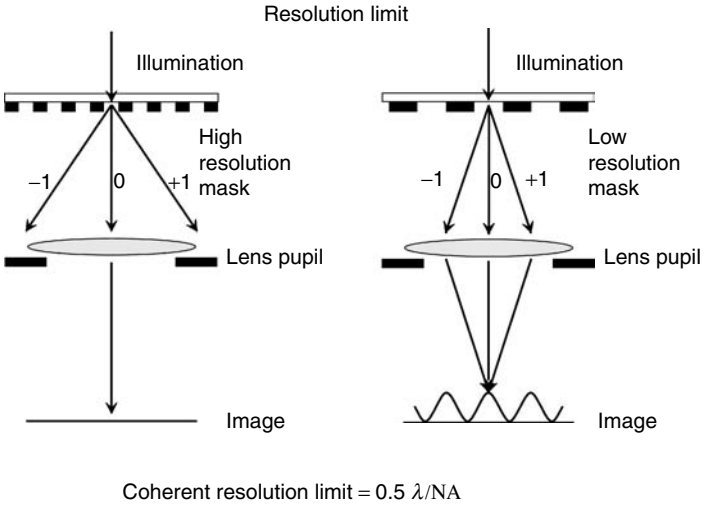


FIGURE 18.14 Schematic mechanism of resolution limit for projection optical system.

The Rayleigh equation can be further extended [50] to include off-axis illumination by adding an additional term,

$$k_1(\text{limit}) = \frac{1}{2\left(\sigma + 1 + \left(\frac{\sin\theta}{\text{NA}}\right)\right)}, \tag{18.19}$$

where θ is the average angle of incidence of the illumination on the reticle. As noted in the discussion above, one must be cautious in applying this sort of analysis. These simple Rayleigh equations are not rigorously correct, and they should be used only for guidance on the effects of various resolution techniques.

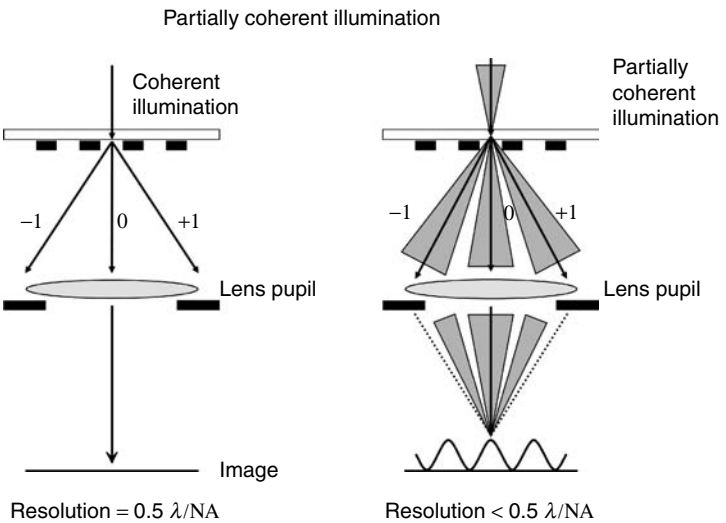


FIGURE 18.15 Mechanism of partial coherence in modifying resolution and imaging characteristics of projection optical system.

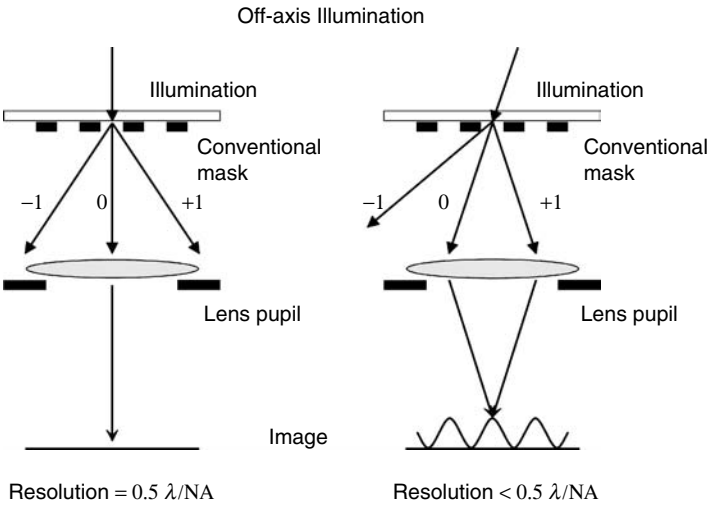


FIGURE 18.16 Mechanism of off-axis illumination for improvement of resolution capability.

18.5.3 Optical Proximity Effect

High-performance optical projection imaging for lithography is strongly impacted by diffraction effects, as noted in several previous sections. One result of this behavior is that individual pattern features do not image independently, but rather they interact with neighboring pattern features. A detailed analysis of the projection imaging process, for example, the analysis described in the paper by Hopkins [51], considers contributions from every portion of the reticle object and every portion of the projection optics in determining the exact image at the wafer plane. A simple heuristic argument considers the extended diffraction structure of the Airy function. Overlap of the diffraction peaks with adjacent pattern features leads to increased or decreased exposure intensity at any point in the image, compared to a purely geometrical image model.

A key result from the diffraction overlap is the so-called proximity effect, in which the exact size of pattern features in the image depends on their proximity to other pattern features. Figure 18.17 shows a simple example. In this case, the lines are exactly the same size on the reticle, but the extended isolated

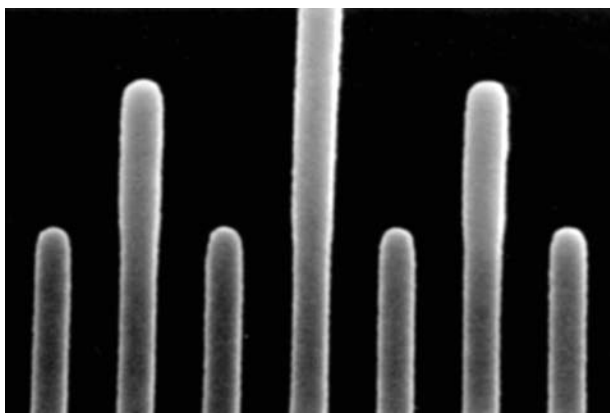


FIGURE 18.17 Optical proximity effect. All pattern features are the same size on the reticle, but the final image is larger for the isolated line extension than for the densely packed lines.

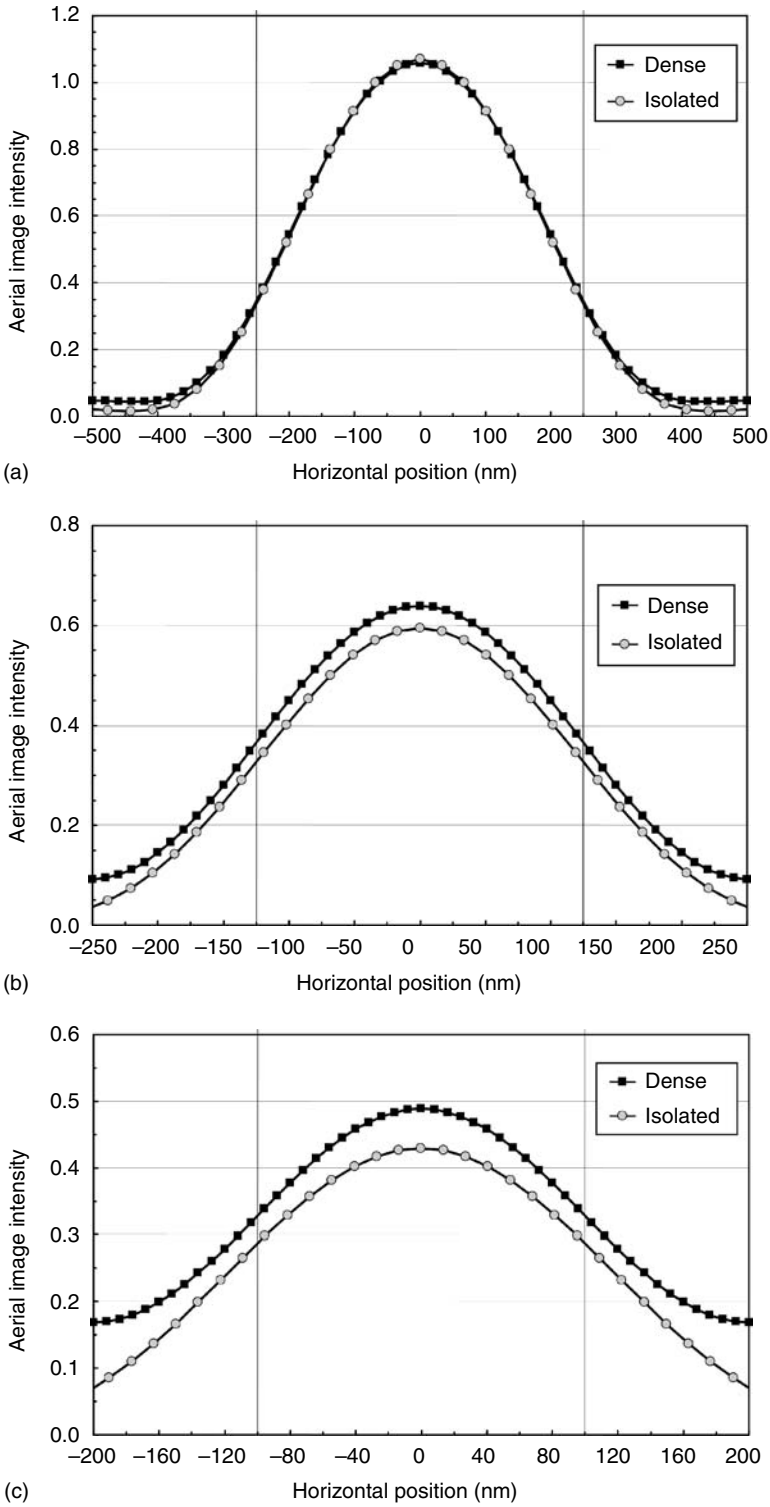


FIGURE 18.18 Aerial image comparison of isolated and dense lines showing optical proximity effect. Exposure with 0.6 numerical aperture (NA) i-line tool. (a) 500 nm lines, (b) 250 nm lines, and (c) 200 nm lines.

line has a larger image on the wafer than the remaining dense lines. A comparison of the aerial image intensity for this case is shown in Figure 18.18. The proximity effect can lead to isolated structures imaging either larger or smaller than densely packed features, depending on the configuration of the imaging system, the resist system, the mask bias, and other factors. As might be expected, the proximity effect becomes much more prominent as the feature sizes and the spaces between the feature sizes approach the resolution limits of the projection optics [52]. In simple terms, the proximity effects become more important as the k_1 factor is reduced below about 0.8.

A typical mask level in circuit layout has a complete range line and space sizes, from the densest possible packing allowed by the lithography technology and the circuit design rules to very sparse packing. In the sparse packing case, the pattern features are typically designated as isolated. Of course circuit layouts consume no more area than necessary, so there is a limit to the degree of isolation of pattern features. However, in practice a pattern feature more than 5–10× its own size from adjacent features can be considered isolated. A common description of the qualitative nature of the proximity effect is shown in Figure 18.19, in which the image width of a specific size line is plotted against the sum of the linewidth and the distance to the nearest neighbor. This sum of the linewidth and the space is usually called the pitch of the line-space pair, even when the line and space arrangement is not periodic.

If the proximity effect is left uncompensated there will be unacceptably large variations in critical linewidths, especially for the CMOS gate level pattern. Therefore, a large amount of research and development has taken place to create effective methods for compensating and preventing the variations. The resulting optical proximity correction (OPC) methods now in use include several pattern adjustment techniques as well as two major computational approaches. The key concepts are outlined below. A sketch of the pattern adjustment techniques is shown in Figure 18.20.

18.5.3.1 Optical Proximity Correction Techniques

Simple biasing consists of an adjustment of the linewidth on the reticle to counter the predicted size variation. Biasing has been used historically for many years to assist in lithography process centering and optimization, but until recently the same bias was applied to the entire reticle. The OPC bias is custom tailored for each line in its own environment.

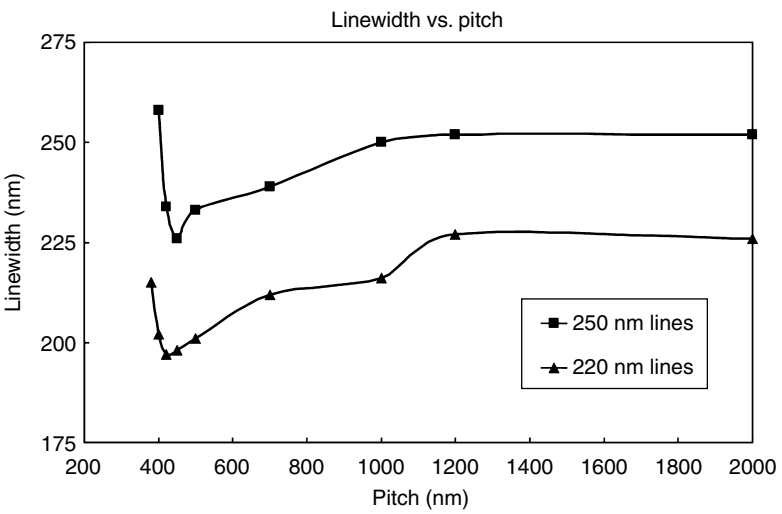


FIGURE 18.19 Optical proximity effect. The imaged linewidth for a fixed reticle pattern size is a function of the space between adjacent lines. Exposures performed with 0.6 NA, 248 nm stepper.

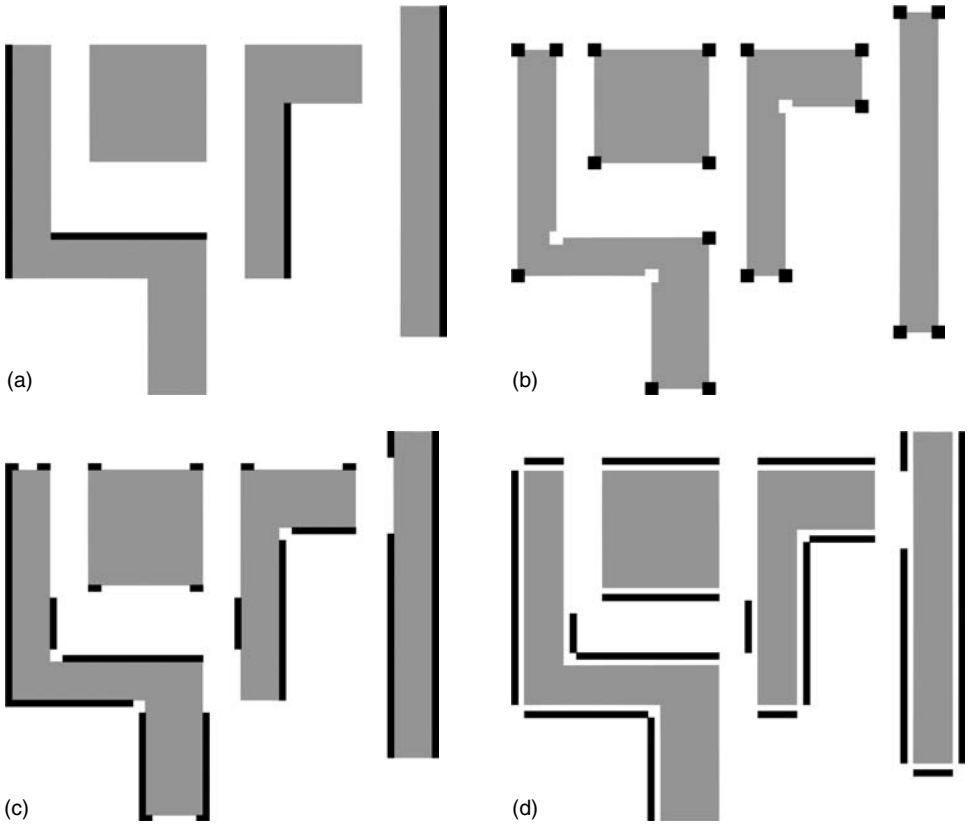


FIGURE 18.20 Adjustment techniques for correcting optical proximity effect. Gray shading indicates the desired pattern, and added features are shown in black. (a) Simple biasing, (b) added serifs, (c) complex biasing, and (d) assistant pattern features to make isolated lines appear dense.

Small pattern features typically called serifs or hammerheads can be combined with the corners or the ends of the main pattern structures to minimize the rounding and end pullback that would otherwise occur. As needed, the extra features can be added or subtracted from the main pattern to give the desired result.

Separate auxiliary pattern features can be placed near isolated lines in attempt to create a dense environment around the isolated lines. The auxiliary features are very small, and they are intended to be below the resolution capability of the projection optics.

18.5.3.2 Optical Proximity Correction Computation

Two primary approaches to applying OPCs to circuit layouts have evolved. The so-called rules-based OPC method combines a predetermined set of rules with a pattern layout database analysis program. The database program systematically determines the proximity environment of every pattern feature edge and then uses a table lookup algorithm to apply the appropriate rules [53]. These rules can generate biasing, serifs, auxiliary patterns, and any other adjustments desired. The database engine is often a standard capability in the circuit design process, and it can also be used for other purposes such as extracting and combining pattern layers, calculating circuit capacitance and inductance, and so on. There are also specialized programs from OPC software suppliers that handle the database efficiently.

The rules may come from several sources. In some cases, they are developed directly from experience in the wafer fab. In other cases, they are developed by image and resist simulation. The most critical factor in the success of rules-based OPC is the selection of an appropriate set of rules. In general, more exact proximity effect correction requires more rules to deal with specific proximity environments.

The other major software pattern correction method is the so-called model-based OPC approach. In this case, each feature and its proximity environment are modeled by an image analysis program. A set of model parameters is used within the program to determine the appropriate corrections to be applied to the pattern layout. The model parameters are typically determined by analysis of specific test structures patterned and processed in the wafer fab. Increasing the correction precision for model-based OPC generally requires a modest increase in the number of model parameters.

There is a trade-off between the rules-based and the model-based OPC methods. The rules-based method is computationally very straightforward, and it in general is quite fast, requiring computation times on typical workstations of a few minutes to a few hours per pattern level. The rule generation is a critical determinant for the success of this method. Generating and validating an extensive set of rules can be a lengthy and tedious process. The rules-based technique is not self-extendable. Corrections will only be applied to a specific pattern if an applicable rule is available.

Model-based OPC is computationally intensive and may require many hours or even days of processing time per pattern level. The model parameters are very important, but there are a limited number required, and the generation is through a well-specified process. The key advantages to the model-based approach are that the correction accuracy and precision can be readily controlled, and the corrections can be successfully applied to any sort of pattern layout. As the lithography requirements have become more stringent due to decreasing feature sizes, the use of model-based OPC is increasing while the use of simple rules-based OPC is diminishing.

In summary, the benefits and limitations of each method can be simplified to the following:

Rules-based optical proximity correction (OPC)	Fast computation	Accuracy and precision completely determined by extent and quality of rules
Model-based OPC	Slow computation	Accuracy determined by quality of model parameters. Precision determined by program settings

There are some important factors common to both OPC methods. The most pressing issue at this time is the significant increase in pattern database size resulting from the OPC process. Pattern databases are defined by rectangles, trapezoids, or turnpoints in more complex polygons. Most implementations of OPC, other than simple biasing, generate many additional pattern features and turnpoints. In the more aggressive correction cases, the increase in the pattern database size can be a factor of 10 or greater. This creates additional difficulty in handling the data, both in the final circuit layout stages and in the mask manufacturing process.

The models or rules created for OPC are tied to a specific lithography process. The important process details encompass the exposure tool parameters such as wavelength, illumination conditions, and NA, and all of the resist parameters such as type, thickness, bake conditions, and developer conditions. In short, anything that can affect the outcome of the lithography process will impact the OPC parameters. Therefore, any significant changes made in the process may require regeneration of the OPC models and rules and may require new reticles.

Finally, it must be recognized that OPC is intended to correct for proximity effects. It does not actually improve the resolution or other quality metrics of the lithography process. It is therefore essential to have a fully capable process even before OPC is applied.

18.6 Manufacturing Considerations

18.6.1 Limits to Optical Lithography

The competitive pressures in the semiconductor industry continue to drive very rapid progress in resolution scaling and in the related attributes of pattern overlay, defect density, throughput, field size, and cost of ownership. It is important to consider not only the key technical attributes, such as resolution, but also to consider the practical manufacturing requirements.

It is apparent from examination of the simple Rayleigh equations that scaling benefits from smaller k_1 factors, higher NAs, and shorter exposure wavelengths. What are the practical limits? [54]

It is notoriously difficult to predict the future of lithography beyond the next few years. Based on an analysis of various reports, press announcements, news stories, and technical articles written in the past 20 years, it has been observed that the end of optical lithography is always about 8 years in the future [55]. These predictions have been made with limited understanding of the practicality of shorter wavelength exposures, higher NAs, phase shift masks, high-performance resists, and other enhancements. An example of the progress in pushing optical lithography beyond initial expectations is shown in Figure 18.21. The advertised and guaranteed performance from the leading exposure tool manufacturers is shown in units of the equivalent k_1 value as a function of year of tool introduction. As can be seen the relative performance of exposure tools has more than doubled in the past 20 years, even without consideration of higher NA and reduced wavelength.

18.6.2 Process Latitude

It is important to maximize the depth of focus and exposure latitude to create a robust process capability. A key determinant for calculating and optimizing the depth of focus is understanding the criteria for measurement. It is customary to consider CD control of $\pm 10\%$ to be a requirement for a high-performance process. In addition to the numerous imaging factors, such as wavelength, NA, and illumination conditions, it is very important to understand the image quality required by the resist. This image quality is typically expressed as contrast [56]. An example of the relationship between contrast requirement and overall system depth of focus is shown in Figure 18.22.

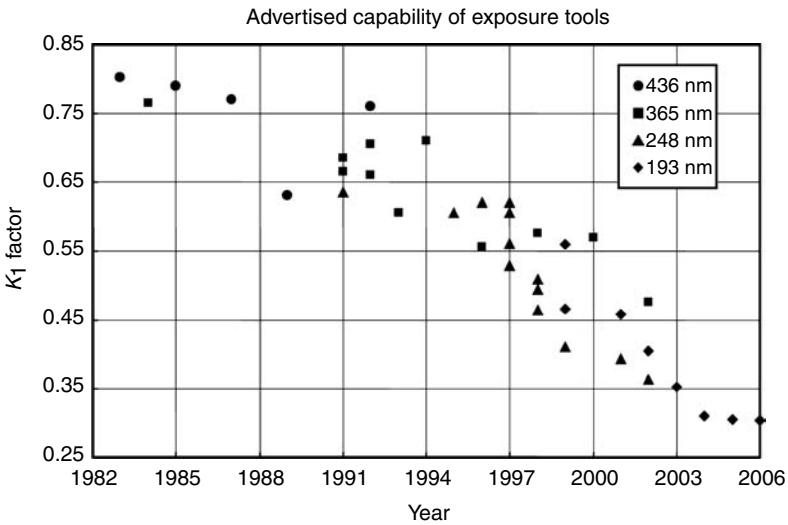


FIGURE 18.21 Advertised capability of exposure tools showing performance improvement through reduction of k_1 in addition to wavelength reduction and NA increase.

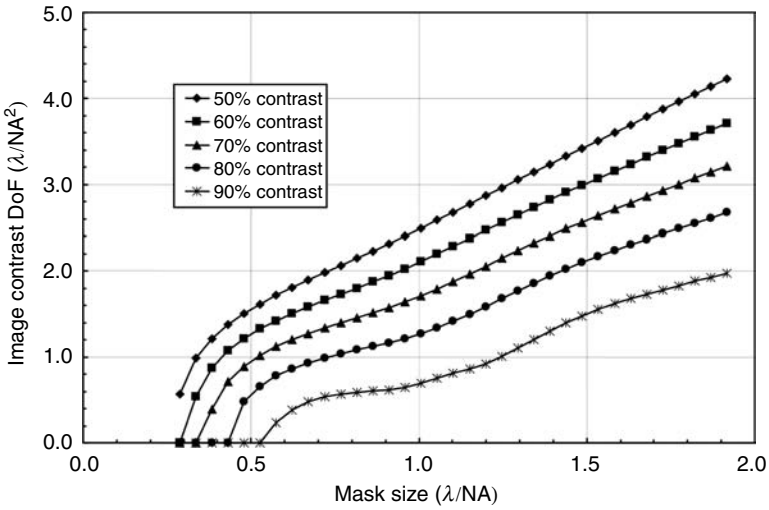


FIGURE 18.22 Working depth of focus as a function of image contrast requirements.

It can be readily observed that selecting a resist process that is able to work effectively with lower contrast images can significantly improve both the resolution and depth of focus. As noted in other chapters, there have been substantial improvements in the contrast behavior of conventional single layer resists, and the use of advanced multilayer resists with thin imaging layers can provide even more capability to use low-contrast images effectively.

Numerous factors impact the required focus budget provided from the imaging system and resist process. These can be grouped into three categories as follow:

- Flatness
  - Wafer substrate
  - Wafer chuck
  - Wafer topography
- Imaging system
  - Image field flatness
  - Astigmatism
  - Focus precision
  - Focus repeatability
  - Image plane leveling
- Process capability
  - Resist thickness
  - Resist contrast.

As outlined in Figure 18.2, the depth of focus (DoF) shrinks rapidly with improvements in image resolution. At a fixed wavelength the DoF scales inversely with NA^2, and even when the wavelength is reduced to improve resolution the DoF scales inversely with NA.

A significant amount of effort is required to optimize the numerous process variables to achieve the best manufacturing capability possible. This is one of the key functions of practicing lithography engineers, and a great amount of research has been conducted and published in this area. There are also several commercial software packages that have been introduced to assist in this optimization [57]. In addition to DoF it is also necessary to maintain a controlled, preferably small, sizing bias between isolated and densely packed features. An example of the type of analysis that is typically performed is shown in Figure 18.23.

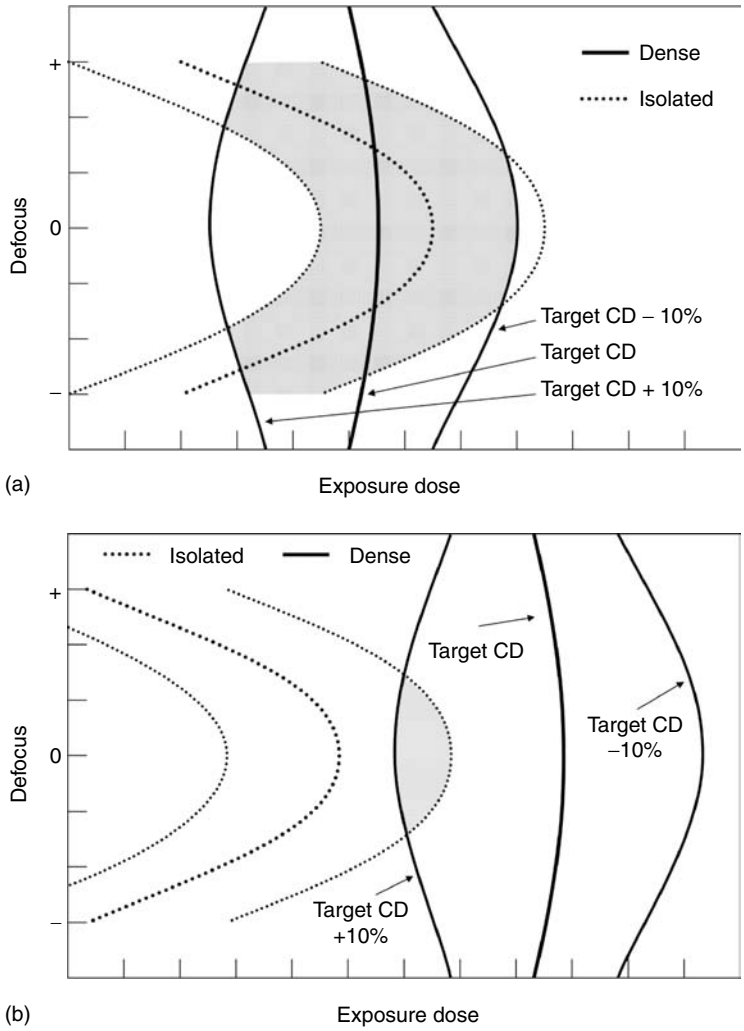


FIGURE 18.23 Exposure–defocus analysis. (a) Isolated and dense lines exhibit large overlap of critical dimension (CD) vs. exposure and focus, giving good process margins. (b) Isolated and dense lines exhibit small overlap of CDs, giving poor process margins.

18.6.3 Mask Error Factor

An important attribute of high-quality lithographic imaging tools and resist processes is linearity. Imaged patterns on the wafer should be the same size as the original patterns on the reticle, after the reduction factor is applied. This is generally found to be true within a few percent for feature sizes down to about 150% of the exposure wavelength. For smaller features, a requirement for leading edge manufacturing, the image transfer from reticle to wafer no longer follows this linear behavior [58]. An example of this non-linearity is shown in Figure 18.24.

This non-linearity effect has been seen for many years, but recently the non-linearity has become very important from a practical point of view. The quality of exposure tool lenses has improved and resolution enhancement technology has provided the means to push quality imaging performance down to and well beyond the wavelength of the exposure light into the regime where non-linearity becomes significant. The increased interest and attention to this phenomenon was spurred by a paper from Maurer on process

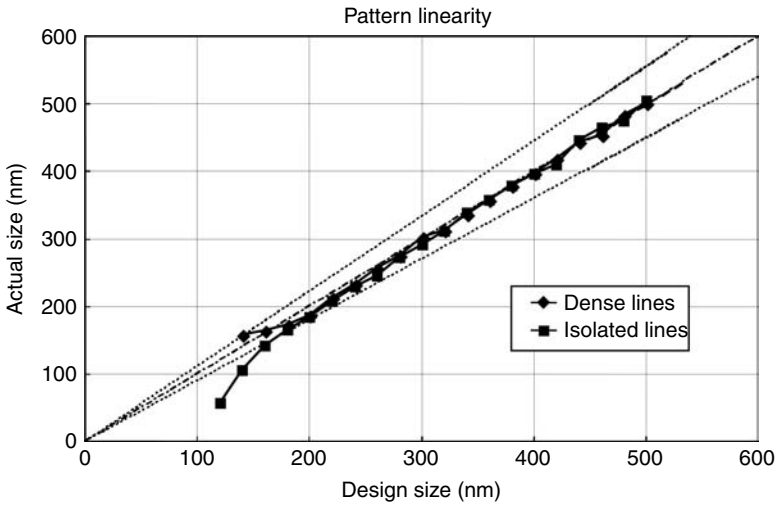


FIGURE 18.24 Linearity of resist size as a function of reticle size for isolated and dense features. Non-linearity is reflected in mask error factor.

optimization for 0.25 μm patterning [59]. The terms "mask error factor" and "mask error enhancement factor" have been applied to this non-linearity effect by later authors. Denoted as either MEF or MEEF, the effect shows itself as an apparent magnification of any reticle CD errors when printed on the wafer. For example, a 30 nm error on the reticle would be expected to result in a 6 nm wafer error for a 5× reduction tool or a 7.5 nm error for a 4× reduction tool. However, some reports show the actual CD error on the wafer to be greater than 20 nm [60–62]. The expected MEF can be calculated directly by using standard lithography simulation tools, such as Prolith or Solid-C.

A study by Wong [60] shows some of the characteristics of the MEF that are now important. By definition,

$$\text{MEF} = \frac{\Delta \text{CD}_{\text{resist}}}{\Delta \text{CD}_{\text{mask}}} \times M \tag{18.20}$$

where M is the nominal magnification of the mask compared to the wafer image. The primary factor in driving the MEF to values larger than 1.0 is the degradation of image contrast as the Rayleigh k_1 is pushed below about 0.6. In an imaging system with no aberrations and no defocus, it has been found that the MEF rises to over 4.0 for a k_1 of 0.35. This corresponds to a resolution of 75 nm for a ArF scanner with an NA of 0.92. This result demonstrates that the error in the resist pattern size can actually be expected to be larger than the error measured on the 4× reticle. Clearly, this presents severe challenges to process control and robust processing.

It has been found that exposure dose has little or no effect on the MEF. The MEF is related to the mask size and the imaging system, not the resist size. Therefore, reticle biasing does not really help reduce the MEF, although it may be desired for other reasons. The MEF degrades somewhat with defocus, but the degradation is small for defocus smaller than a Rayleigh unit of defocus, $\lambda/2\text{NA}$.

The largest factor in controlling the MEF is improvement in the image contrast. This can be accomplished in numerous ways, including the use of more advanced exposure tools (shorter wavelength, higher NA, better aberration control), and through the use of various RETs. Of particular benefit, at little cost, is the use of annular or quadrupole illumination. The disadvantage is that the proximity effects tend to be emphasized and the isolated-to-dense feature bias control suffers.

A rigorous mathematical characterization of MEEF, with particular emphasis on application to algorithms for OPC, has been published by Granik [63]. This 2D treatment supports correction for complex patterns such as line ends, corners, and nested structures.

18.6.4 Mix-and-Match Lithography

Lithography remains the single biggest cost component for advanced wafers fabs manufacturing complex VLSI chips. Therefore, it is very important for the lithography systems to operate as efficiently as possible, with the lowest total cost of ownership that can be achieved. There has been a continuous progression in the capability of lithography exposure tools over many years. Improvements in resolution and image placement have resulted from advanced mechanical systems, from full field to step-and-scan tools, as well as wavelength reductions. However, along with the technology capability improvements the capital cost of leading edge exposure tools has also increased at a steady rate. A recent analysis shows an exponential increase over many years [64]. It is therefore beneficial to use the most advanced and most expensive tools only where absolutely required and to use lower cost tools where possible. The most critical patterning levels in an advanced CMOS VLSI process are the active area, the polysilicon gate, and the first few levels of interconnect, including both leads and contacts. Less demanding patterns, such as implant levels and the higher interconnect levels, are more cost-effective if they are performed with lower capability tools. As this is written, the so-called mix-and-match strategy employs high NA 193 nm excimer laser scanners for the critical levels, moderate NA KrF scanners for the intermediate levels, and moderate NA i-line steppers or scanners for the less critical levels. Detailed descriptions of cost of ownership models are included in other chapters in this volume. Applications to lithography have been described in the literature [65,66].

The primary challenge to a successful mix-and-match strategy is achieving the required pattern overlay tolerances between the different exposure tools. Three categories of overlay errors are important in this analysis. First, there are residual image placement differences between even tools that are nominally identical. This can drive the lithography engineer to specify that exactly the same tool must be used for active area pattern and polysilicon gate pattern, for example. While not a direct impact to capital cost this requirement creates difficulty in maximizing the overall throughput and output of the wafer fab.

The more typical mix-and-match concern is the overlay of patterns from different types of tools, such as a high NA 193 nm scanner and an i-line stepper. In this case, there will be systematic differences in the residual distortion in the projection lens as well as differences in stages and alignment systems. No attempt will be made here to analyze the error terms in detail. Each case will be different, and the complete product and technology teams must agree on the trade-offs between tool costs, design rules, and product performance.

An additional mix-and-match issue has arisen in the past few years. With the continuing advancement in resolution enhancement, the effective k_1 factor has been reduced to levels approaching 0.4 and even lower. In these cases, the CD control across becomes a very sensitive function of the bias between isolated and dense features and of reticle errors. Small changes in the aberration signature between exposure tools can lead to CD control and uniformity errors due to variations in the isolated-to-dense line bias described in Section 18.5.3. Since the correction for optical proximity effect is designed into the reticle, it is a severe limitation on mix-and-match strategies if nominally identical tools require different reticles. Indeed, in most cases, this would be considered cost prohibitive. Therefore, a key component of a successful mix-and-match strategy is the calibration and tuning of optical proximity effects in the tools to be matched. This calibration and tuning is typically done with adjustment in the illumination and NA of the lenses, as well as directly adjusting the aberration components where possible.

18.6.5 Error Budgets

The two major patterning error categories, CD control and image placement, often are characterized and analyzed in terms of so-called error budgets. These budgets attempt to identify and quantify the sub-components of the overall CD and image placement errors.

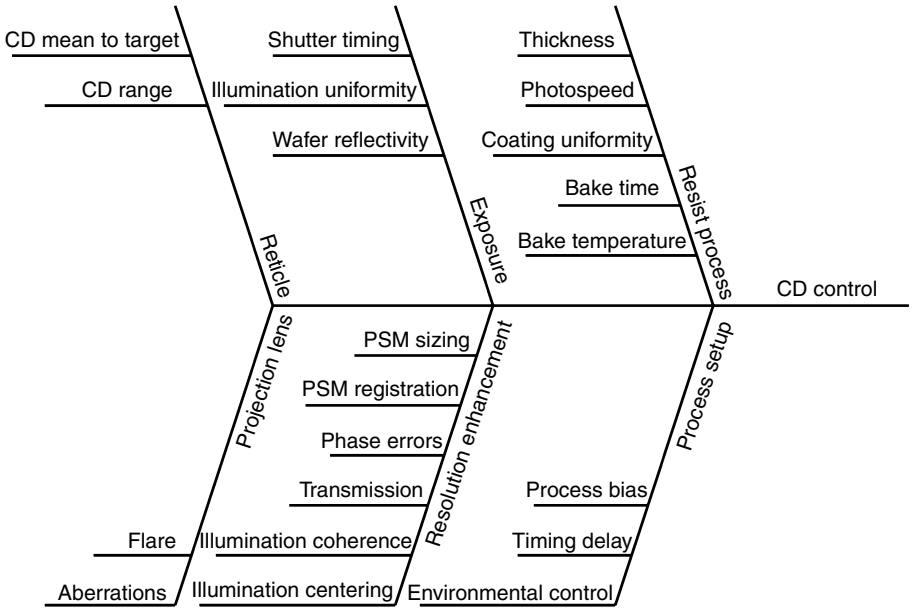


FIGURE 18.25 Critical dimension (CD) control error sources.

Factors that need to be included in a CD control budget can be shown on a fishbone diagram such as Figure 18.25, and the factors in an image placement control budget are shown in Figure 18.26.

The summation of error source components to predict the total error expected is generally performed as a combination of algebraic addition of systematic errors and RSS contributions from random error

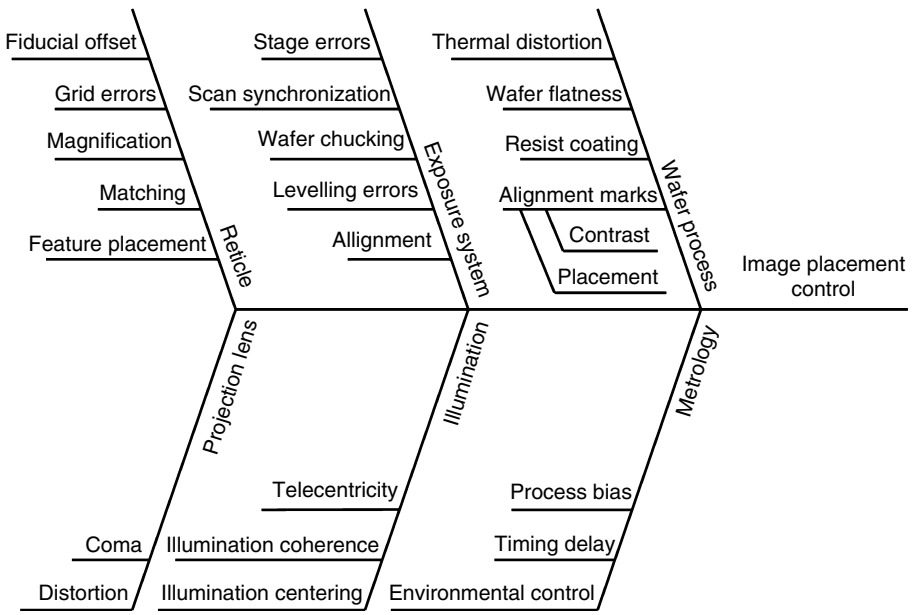


FIGURE 18.26 Image placement (IP) error sources.

terms. Analysis of lithography errors employs the same methods widely used in all technical and scientific disciplines. An important detail is careful definition of what constitutes systematic and random errors. As an example, the placement and CD errors built into a reticle will have both systematic and random components, depending on the exact characteristics of the reticle manufacturing tools and process. However, once in use in the wafer fab, the reticle errors are entirely systematic. The same errors will apply to every exposure on every wafer processed with a given reticle. In other cases, the error may be random over a long period of time, but within the processing time of a single wafer or one lot of wafers the error is constant and systematic. An example of such an error would be a baseline error in the alignment system. This is typically checked on a periodic basis, so over a long time period the average error is corrected. However, over short time periods any drift might be undetected and uncorrected. It is important that each pattern geometry on each chip on each wafer is correct. Acceptable long-term error averages with significant short-term fluctuations may lead to disappointing yields. This characteristic of wafer patterning has led some lithographers to use the so-called good fields rule for analyzing errors rather than traditional addition and RSS combination [27]. It is of course possible and often worthwhile to perform more sophisticated statistical analysis of CD errors and image placement errors. A publication by Wong discusses measurement methods and analysis techniques for advanced characterization of linewidth variation [67].

18.7 Recent Advances in Optical Lithography

18.7.1 Immersion Lithography

In recent years, the progress in resolution and image placement from optical lithography tools has been continuous and rapid. At the time of this edition, there is widespread manufacturing use of exposure tools working at the 193 nm wavelength of the ArF excimer laser. These tools have NA greater than 0.9, providing useful line/space resolution capability down to about 65 nm. As shown in Figure 18.2, however, it is not possible to increase the NA much farther if the exposure takes place in air. As noted previously, attempts to extend the resolution capability of optical lithography by moving to even shorter wavelengths, namely 157 nm, were deemed to be too difficult and too costly to pursue toward full manufacturing implementation. At the same time, the process factor, k_1, in the resolution Equation 18.3 has been reduced through improved processes, reticles, and exposure tools to a value approaching the theoretical limit.

The remaining factor, NA, can be increased beyond 1.0 if the exposure medium (air) is replaced by a material with an index of refraction greater than 1.0. It has been the practice for many years to increase the resolution of microscopes by filling the space between the bottom of the lens and the object to be imaged with a transparent oil. This technique has recently been extended to optical lithography by the addition of water between the exposure tool lens and the wafer [68–70]. Figure 18.27 shows the basic imaging behavior for conventional exposure in air and for immersion lithography using water. The NA for air exposure is limited to 1.0. Any attempt to increase the optical ray angles further, thereby increasing NA, would simply lead to total internal reflection of the light back into the lens. Introduction of the exposure medium, in this case water, allows the rays to pass on to the resist. It should be noted that the optical ray angles in the resist are not affected by the imaging medium. Therefore, there is no direct impact on resolution capability by the addition of the immersion fluid, as long as the NA remains below 1.0. There may be secondary effects, including an improvement in DoF as discussed below. The immersion fluid does open up the possibility of NA greater than 1.0, and this is where the real benefits lie.

The basic imaging properties of this so-called immersion lithography are understood from a straightforward modification of the standard resolution and DoF equations shown in Figure 18.2. This figure has been redrawn here as Figure 18.28. Equation 18.7 is updated to:

$$\text{Resolution} = k_1 \frac{\lambda}{n \sin\theta} = k_1 \frac{\lambda}{n} \frac{1}{\sin\theta}. \tag{18.21}$$

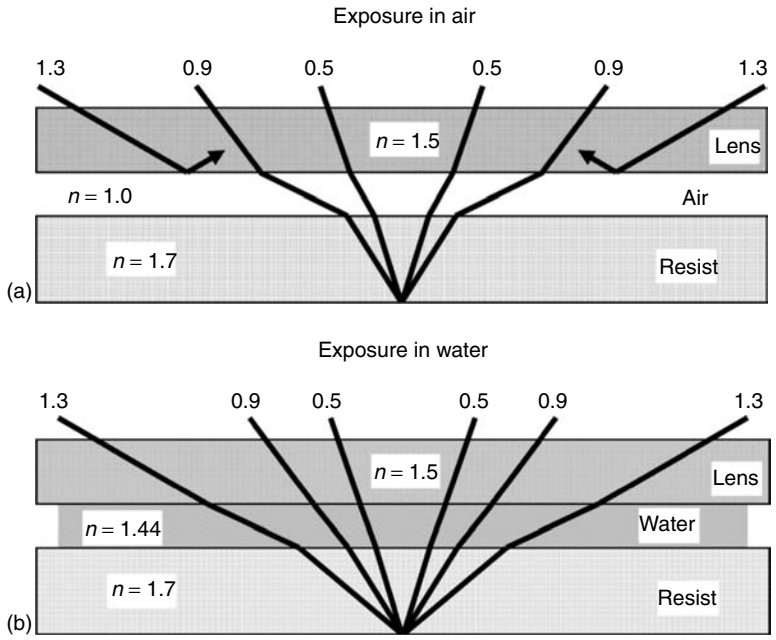


FIGURE 18.27 Optical ray trace sketches showing increase in achievable NA provided by immersion exposure number above each ray is the NA represented by that ray.

This expression shows that immersion lithography has the same effect as reducing the exposure wavelength by the refractive index of the immersion fluid. In the case of pure water, the index at 193 nm is about 1.44, and the effective exposure wavelength is 134 nm. One must be careful not to take this alternate analysis too far. For most purposes, the wavelength must still be taken as 193 nm. It is merely a demonstration of the possible resolution improvement that comes from immersion lithography.

The Rayleigh equation for DoF, shown in Figure 18.2, is derived from considerations of the angle of the extreme rays in the medium between the lens and the wafer. In this case, the effective NA in that medium is *reduced* from the system NA by a factor of n, the refractive index of the medium. The Rayleigh DoF equation can be extended to:

$$\text{DoF} = k_2 \frac{\lambda}{n} \frac{1}{(\text{NA}/n)^2} = nk_2 \frac{\lambda}{(\text{NA})^2}. \tag{18.22}$$

We can see from this elementary analysis that the DoF provided by immersion lithography can be expected to increase from the DoF available from dry (air between the lens and wafer) lithography by a factor of n. Detailed optical analysis shows that the increase of DoF can be expected to be even greater than n [5]. This has been verified by experiment for NA less than 1.0. DoF improvement of up to a factor of 2 has been observed. Exposure in air is not possible for NA greater than 1.0, but the expected improvement in DoF is seen from early experiments with so-called hyper-NA immersion exposure tools [71–73].

The potential benefits of immersion lithography are clear; extended resolution capability and greater depth of focus are unquestioned. Production-ready exposure tools employing water as the immersion fluid have entered the market. Semiconductor manufacturers have strongly encouraged and supported the introduction of immersion technology [74]. However, there are also added challenges that come with the use of immersion. The most obvious is the need to establish and control the layer of water between the bottom of the projection lens and the wafer. There are at least three conceivable

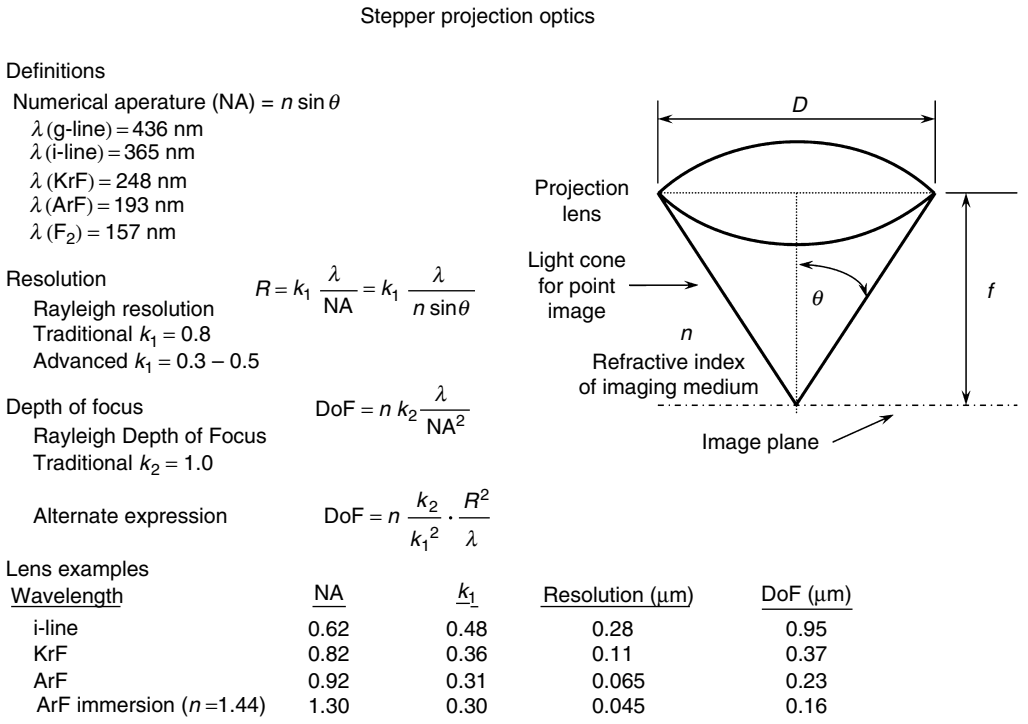

Stepper projection optics
Definitions
Numerical aperature (NA) = $n \sin\theta$
λ (g-line) = 436 nm
λ (i-line) = 365 nm
λ (KrF) = 248 nm
λ (ArF) = 193 nm
λ (F_2) = 157 nm
Resolution
$R = k_1 \frac{\lambda}{NA} = k_1 \frac{\lambda}{n \sin\theta}$
Rayleigh resolution
Traditional k_1 = 0.8
Advanced k_1 = 0.3 – 0.5
Depth of focus
$DoF = n\, k_2 \frac{\lambda}{NA^2}$
Rayleigh Depth of Focus
Traditional k_2 = 1.0
Alternate expression
$DoF = n \frac{k_2}{k_1^2} \cdot \frac{R^2}{\lambda}$
D
Projection lens
Light cone for point image
θ
f
n
Refractive index of imaging medium
Image plane
Lens examples
Wavelength | NA | k_1 | Resolution (μm) | DoF (μm)
i-line | 0.62 | 0.48 | 0.28 | 0.95
KrF | 0.82 | 0.36 | 0.11 | 0.37
ArF | 0.92 | 0.31 | 0.065 | 0.23
ArF immersion (n =1.44) | 1.30 | 0.30 | 0.045 | 0.16

FIGURE 18.28 Basic relationships in projection optical systems. Updated to include immersion.

approaches. One would involve complete immersion of the wafer, the wafer stage, and the bottom of the lens. Although this might work for imaging purposes, the mechanical difficulties of accommodating the rapid stepping and scanning motions required for complete imaging of the entire wafer would be very large. There are no reports of attempts to develop an exposure system using this method. A second method would limit the water to a smaller region fully contained in the wafer stage. A small pool would move with the stage, and there would be no need for the stage itself to travel through a large bath of water [75]. Again, there are no reports of usage of this approach. The third approach, which has been adopted by all of the major exposure tool manufacturers, creates a localized film of water contained between the bottom of the lens and the wafer. This film is essentially stationary under the lens, and the necessary stepping and scanning motions are accomplished by moving the wafer under the film. The film is constantly refreshed by the flow of water from a fill port to a removal port [76,77].

The key challenges for practical implementation of water-based immersion lithography include:

- optical design of projection lens,
- mechanical system design to maintain high productivity and high precision,
- additional wafer defects due to immersion,
- control of the water-resist surface effects, and
- thermal control of the water.

Each of these will be discussed briefly.

18.7.1.1 Optical Design of Projection Lens

Initial exposure tools for immersion lithography are based on extensions of the fully refractive lens designs that have been used for ordinary dry lithography. In this manner, it is possible to develop exposure tools with NA approaching 1.10 while still maintaining a standard exposure field scan width of

26 mm. It is theoretically possible to extend these refractive designs to even higher NA, but the practical complexities increase very rapidly. Larger NA greatly increases the required size of the individual lens elements, impacting both cost and manufacturing difficulties [78]. The alternative choice is to use a catadioptric lens design, employing a small number of reflective elements in addition to refractive elements. There are numerous design possibilities, but in general the reflective elements provide the imaging power of the lens while the refractive elements provide for the correction of aberrations. Catadioptric designs typically have three or four reflective elements (mirrors), and the architecture can be in-line or multi-axis. It is possible to increase the NA up to 1.30 or greater, while maintaining overall lens size and complexity that is similar to the largest possible all-refractive designs [79].

A sketch of one form of catadioptric projection lens is shown in Figure 18.29. In this example, which is a generic representation only, three mirrors are employed. Two mirrors are simple folding mirrors, shown here as two faces on a single substrate, and the third (curved) mirror provides a substantial portion of the imaging power of the lens. The various refractive lens elements, which again are only representative, serve as correctors for the imaging aberrations in addition to providing some of the imaging power of the lens.

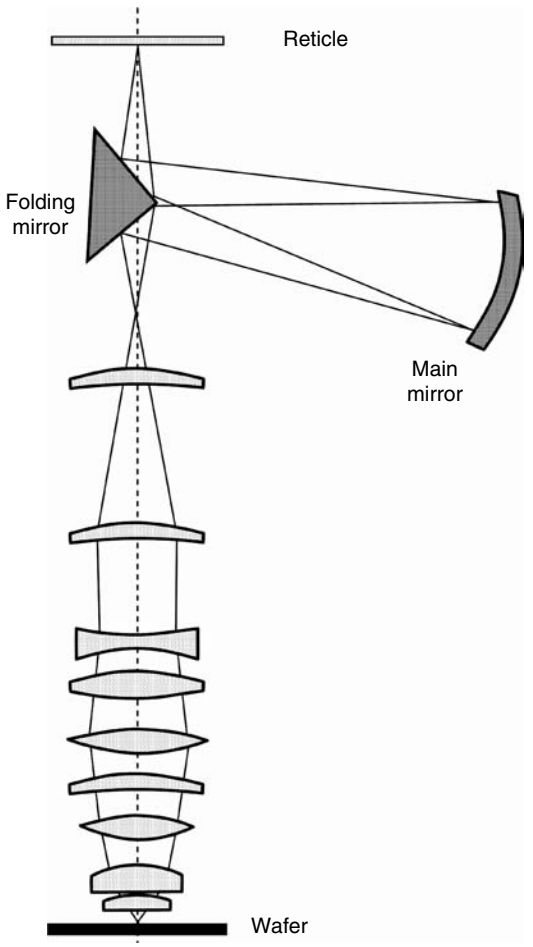


FIGURE 18.29 Generic example of one form of advanced catadioptric projection lens.

18.7.1.2 Mechanical System Design to Maintain High Productivity and High Precision

Introduction of a fluid into the space between the lens and wafer introduces several key issues. The fluid (water) must be contained and controlled so that it remains where needed and does not escape to cause damage to the exposure tool and wafer. The thin film of water, generally no more than about 1 mm thick, is maintained in position primarily by surface tension properties of the nozzle materials. In addition, some implementations add an air curtain to help keep the water in place [73]. In order to maintain high productivity and high precision, it is essential that the wafer stage motion is not degraded by the introduction of the immersion fluid. The scanning speeds must remain high, and image overlay precision requirements become even more stringent when the image resolution is improved. There is now a substantial increase in shear force between the lens and the wafer, but this can be accommodated by modern measurement and control systems used in all exposure tools.

There are numerous practical issues with startup and shutdown, focus, and position metrology, and wafer exchange. These are tool-specific details beyond the scope of this chapter, and they will not be further discussed.

18.7.1.3 Additional Wafer Defects due to Immersion

One of the most critical remaining issues for immersion lithography at the present time is the control of defects. The addition of water into the exposure process creates the potential for several new categories of yield-damaging defects. First is the possible existence of air bubbles or floating particles in the water itself. Each of these could block or scatter the imaging light from the lens. Particles, of course, must be completely avoided by filtration and standard techniques for handling high-purity water in the fab. In this sense, there is little difference from other fab processes. Bubble impacts are unique to immersion lithography. Larger bubbles, say 1 μm in diameter or larger, will simply block the light in the same manner as a particle. These must be completely avoided. Smaller bubbles, on the order or 100 nm or less, will have little impact unless they are close to the wafer surface. In that case, the impact can range from complete blocking of the light to subtle distortion of the patterned image on the wafer [80]. Again, avoiding the small bubbles is important. However, these are naturally more common, and great care must be given to the immersion nozzle design and details of the water flow.

An important class of potential defects comes from the interaction of the water with the resist. It has been shown that leaching of the chemical components of typical 193 nm resists occurs rapidly when the resist is contacted by water. This in turns can lead to a change in resist imaging sensitivity or in some cases complete failure of the resist patterning. Several possible solutions have been pursued. The resist formulations can be modified to give more stable performance in a water environment. A barrier or topcoat can be added to the top of the resist to prevent or at least greatly reduce the interaction of the water and the resist. Added process steps such as pre- and post-soaking the resist can create more uniform resist behavior over the entire wafer. This prevents imaging failures related to pattern density differences across the wafer or differences in the exact amount of immersion time at every point on the wafer. A negative consequence of both the topcoat and the soak procedure is the extra process steps that are required for the complete lithography process. In general, additional modules are required on the resist processing track.

18.7.1.4 Control of the Water-Resist Surface Effects

The configuration of the water containment system in practical immersion lithography exposure tools is based on some sort of nozzle located around the periphery of the exposure field at the bottom of the lens. As noted previously, the actual water containment is primarily a function of careful control of the surface tension properties of the nozzle materials. It is equally important to control the surface interactions between the water and the resist-covered wafer. For ease of maintaining control of the water film, it is most favorable that the resist surface is hydrophobic. In this case, there is little tendency for the resist to drag water from the containment area. On the other hand, a hydrophilic surface will tend to pull out either droplets or a complete water film as the wafer stage moves under the lens. It has been learned both from experiment and analysis that the optimum condition of the resist surface occurs when the contact

angle of water on the resist is 70° or higher. This degree of hydrophobicity is adequate to allow containment of the water film under the lens [70].

Ordinary 193 nm photoresist used for dry lithography has no such requirements on surface conditions, at least for the exposure part of the lithography process. Many conventional resists have marginal or inadequate contact angles for use in immersion lithography. These resists need to be modified or they need to be covered with a hydrophobic topcoat material. As noted previously, the topcoat requires added process steps. However, the use of a topcoat allows the full optimization of the both the imaging performance and the mechanical performance of the resist.

18.7.1.5 Thermal Control of the Water

Temperature control is vital in any modern exposure tool. Optical and mechanical precision requirements in the nanometer regime require close attention to any effects of thermal expansion of materials in the positioning systems or in the optical properties of the materials in the image path. When water is introduced into the critical position between the lens and wafer it creates the need for even more careful control. The water introduces two new complications. First, the heat capacity of the water is much larger than air, so it is essential that the immersion water flow temperature is maintained to within a few milliKelvin of the target temperature [71]. This requirement is driven both by the possible wafer size change with a temperature change and the change in the index of refraction of the water with a temperature change. Temperature variations can have a significant impact on the pattern overlay precision due to the thermal expansion coefficient of silicon. Any change in the refractive index of the water beyond a part per million will lead to degraded imaging. Since, the change of water refractive index ($\Delta n/n$) is about 10^{-4}/K [81] it is necessary to control the water temperature to less than 10 mK.

The second concern with temperature control comes from the cooling that results when there is any evaporation of the water. It is essential that all of the water introduced into the flow is recovered and that none of it evaporates. This effect can be particularly troublesome if the cooling is uneven across the wafer. Such non-uniformity can arise from either evaporation in the water containment process in the immersion nozzle or in residual water films and water droplets remaining on the wafer.

18.7.2 Polarization

In the earlier stages of optical lithography, when the lens NA was smaller than about 0.5 it was typical to consider the imaging function as a scalar process. In particular, with the exception of some lens design details there was little or no attention paid to the polarization state of the incident light coming through the reticle and into the projection lens. As the resolution capability of lithography tools increased through the increase in the lens NA, it was widely understood that polarization was beginning to have observable impact on the imaging performance. However, impact was still quite small, and the exposure tools continued to provide only an unpolarized illumination beam onto the reticle and through the projection lens. As the NA increased beyond 0.9, it became important to include flexible polarization control as part of the exposure tool. At the hyper-NA conditions used in immersion lithography, polarization control becomes mandatory for the best imaging performance [82–84].

Why is polarization control important for high NA and high-resolution imaging? Figure 18.30 shows the basic nature of imaging high-resolution linear features by considering the optical ray path from the reticle to the image. The sketch on the left shows the path of the diffracted imaging rays from the reticle to the wafer. The sketches on the right show the orientation of electric field vectors for transverse magnetic (TM) polarization and transverse electric (TE) polarization. In the TE case, the vectors from the two diffracted rays are parallel and upon recombination in the image plane they interfere constructively to form a high-contrast image. This constructive addition occurs for any NA. On the other hand, the electric field vectors in the TM case are not parallel. They constructively combine only partially, and the imaging degrades as the NA increases. The resulting image contrast can be defined by simple geometrical-based equations as shown in Figure 18.31. This figure shows the extremes of complete TE polarization and complete TM polarization. Intermediate cases, including unpolarized illumination, will exhibit image

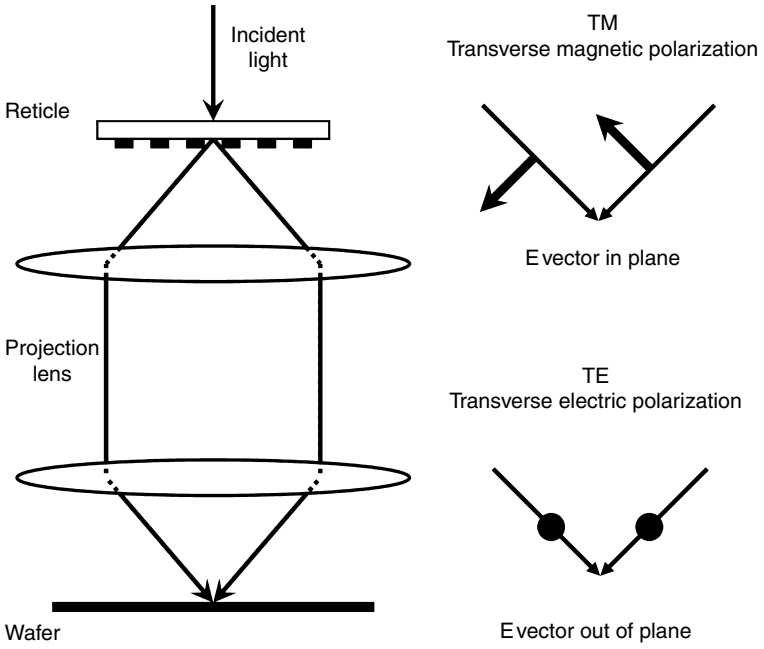


FIGURE 18.30 Image formation from polarized light.

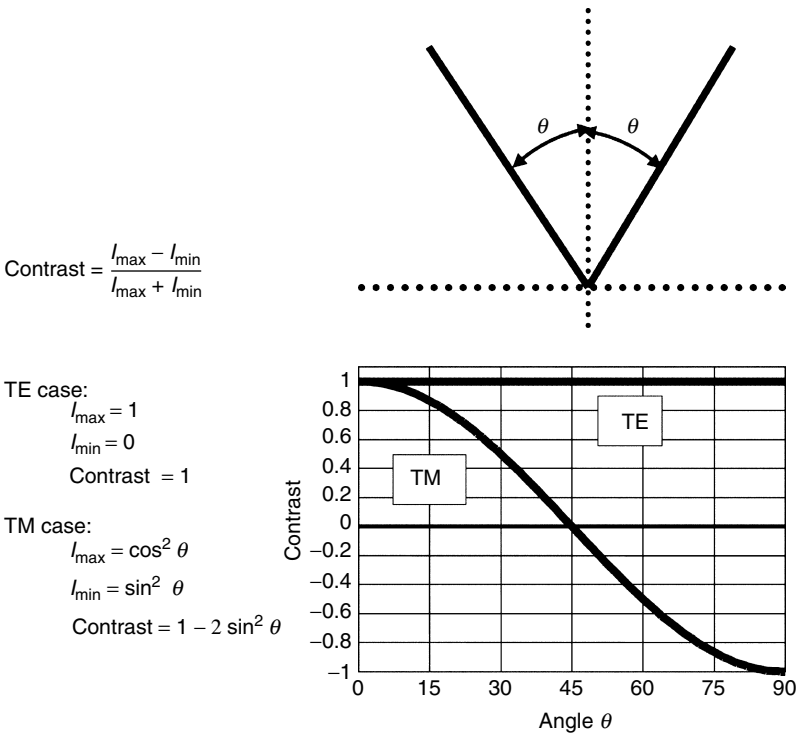


FIGURE 18.31 Image contrast when imaging light is polarized.

contrast between the two extremes. Detailed discussion of high NA imaging and polarization effects is found in several recent publications [84–88].

18.8 Patterning Roadmaps

Lithography has long been regarded as the pacing item in the shrinking of structures on semiconductor wafers. As such, there is a great interest in the trends and forecasts for lithography capability. Many companies and individual researchers have forecasted such capabilities, often with different conclusions. To help reduce confusion and to provide a common reference for semiconductor manufacturers, equipment, and materials suppliers, and other interested parties, the Semiconductor Industry Association undertook a national (U.S.-based) effort in 1992 to create an industry-wide needs and capabilities roadmap. The initial effort was followed by a more comprehensive roadmap, the so-called National Technology Roadmap for Semiconductors in 1994. Comprehensive updates were made in 1997 and 1998. Beginning in 1999 the roadmap included input from around the world, and the name was changed to the ITRS. The ITRS updates have continued on an annual basis.

The key lithography requirements from the 2005 ITRS update are shown in Table 18.3. The use of "Technology Node" as a descriptor has been officially eliminated, but such usage is still very common. It is immediately apparent that the feature sizes required from 2005 and beyond are far smaller than the wavelength of the available optical lithography exposure tools. While there are several alternative lithography systems proposed and under development around the world, it is expected that optical microlithography will prevail as the dominant patterning method until well beyond 2010. Therefore, it is essential to understand and practice optical lithography enhancement techniques to provide semiconductor manufacturing capability until some replacement technique becomes available and cost-effective.

Changes to the roadmap have been quite significant on a year-by-year basis. In particular, the classic view that the leading edge of semiconductor production follows a feature size scaling of 70% every 3 years has become a subject of some debate. This scaling behavior is usually tied to Moore's Law [89,90], which describes the functionality of semiconductor integrated circuits as an increasing straight line on a semi-logarithmic scale. For many years, the integration model followed by dynamic random access memories was a new generation every 3 years, with each generation providing 4× the bit density of the previous generation. The 4× increase was supported by a packing density improvement of 2× due to the area reduction from the 70% feature size scaling (area scaled by 0.7^2), an increase in chip area of 1.4×, and a circuit innovation factor of 1.4×. The innovation factor included trench bit cells, stack cells, and other circuit layout improvements. The scaling rate for feature size, chip area, and even circuit innovation has been debated in recent years. The ITRS timelines have gone from 3- to 2-year cycles and back again. For the latest information, the ITRS should be consulted. As noted throughout this chapter, the resolution

TABLE 18.3 International Technology Roadmap for Semiconductors (ITRS) Product Critical Level Lithography Requirements

Year of First Chip Shipment	2005	2007	2010	2013	2016	2019
Technology node (nm)	80	65	45	32	22	16
Isolated lines (Logic Gates) (nm)	32	25	18	13	9	6
Dense lines (DRAM) (nm)	80	65	45	32	22	16
Dense lines (Logic) (nm)	90	68	45	32	22	16
Dense lines (Flash) (nm)	76	57	40	28	20	14
Contacts (DRAM) (nm)	85	64	45	32	23	16
Contacts (Logic) (nm)	101	77	51	36	25	18
Gate critical dimension (CD) control (3 sigma) (nm)	3.3	2.6	1.9	1.3	0.9	0.7
Overlay (3 sigma) (nm)	15	11	8	5.7	4.0	2.8

capability provided by optical lithography has continued to improve but with increasing complexity and expense. Chip area can continue to increase if desired, but increased cost and yield degradation are strong motivation factors toward keeping the chip scaling as small as possible. At the present time, the 70% scaling every 2–3 years is still in the roadmap forecast, but this rate may change in the next few years.

18.9 Summary

Optical lithography has served as the dominant patterning technology in the semiconductor industry since the integrated circuit was invented nearly 50 years ago. Improvements in resolution, image placement, and pattern information transfer rate have been enormous. This chapter has outlined some of the basic principles of optical lithography and has attempted to describe some of the non-traditional enhancement techniques that are now becoming important for mainstream manufacturing use.

An often asked question is how long optical lithography can continue to be the mainstream patterning technology for the semiconductor industry. Predictions of the end of optical lithography have been made for more than 20 years, typically within 5–10 years from the date of the prediction. Of course, this has not happened, and it is instructive to consider the reasons for the longevity of optical lithography well beyond virtually all predictions. The three primary reasons are (1) Optics were not really up against any fundamental limits. Significant improvements in NA, aberration control, wavelength reduction, and RETs have continued to be made. More recently, it has become feasible to extend the resolution through the use of immersion exposure. (2) The proposed replacement technologies were not ready as predicted, either in technical capability or cost-effectiveness. (3) The infrastructure capability continues to advance. Major improvements in optical materials, lens manufacturing methods, mechanical systems control, and computer integration of exposure tools have allowed error allowances to be reduced and the lithographic performance pushed to nearly theoretical limits. A sketch of the important contributors is shown in Figure 18.32.

While the fundamental physical limits for optical lithography may still be quite far away, it is clear that there are practical limits to NA increase and wavelength reduction, at least for the type of systems in use today. Novel optical approaches have been proposed that could push the capability of optical lithography

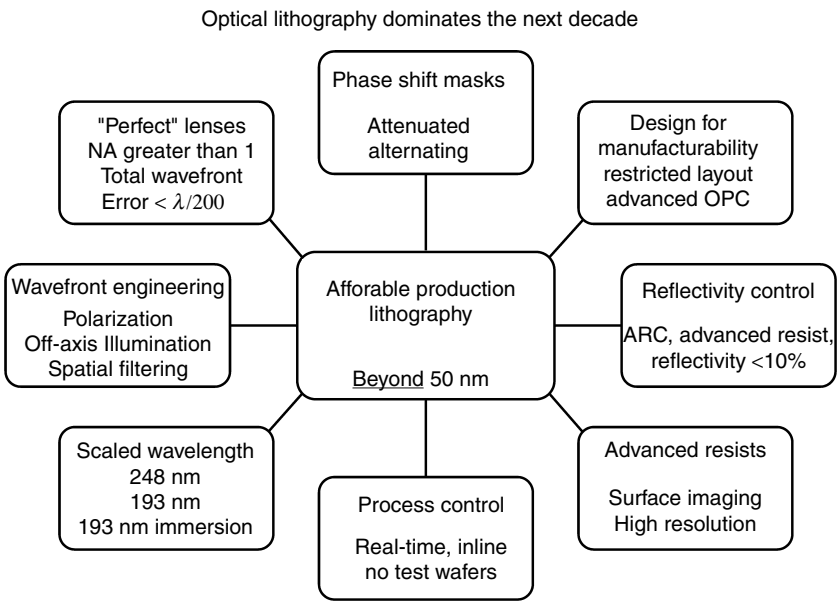


FIGURE 18.32 Outlook for optical lithography.

well beyond simple scaling of the tools that in use today [91]. Even without radical changes in tool design it appears that optical lithography will prevail for at least another 5–10 years, supporting semiconductor technology beyond 50 nm line and space patterning.

References

1. Bruning, J. H. "Optical Lithography—Thirty Years and Three Orders of Magnitude." *Proc. SPIE* 3051 (1997): 14–27.
2. Levinson, H. J., and W. H. Arnold. In *Optical Lithography*, Vol. 1, edited by P. Rai-Choudhury, *Handbook of Microlithography and Microfabrication*, Bellingham, WA: SPIE, 1997.
3. Sheats, J. R., and B. W. Smith, eds. *Microlithography: Science and Technology*. New York: Marcel Dekker, 1998.
4. Bruning, J. H. "Optical Imaging for Microfabrication." *J. Vac. Sci. Technol.* 17, no. 5 (1980): 1147–55.
5. Lin, B. J. "The k_3 Coefficient in Nonparaxial λ/NA Scaling Equations for Resolution, Depth of Focus, and Immersion Lithography." *J. Microlith. Microfab. Microsyst.* 1, no. 1 (2002): 7–12.
6. Airy, G. B. *Trans. Camb. Phil. Soc.* 5 (1835): 283.
7. Rayleigh, L. *Phil. Mag.* 8, no. 5 (1879): 403.
8. Toh, K. K. H., and A. R. Neureuther. "Identifying and Monitoring Effects of Lens Aberrations in Projection Printing." *Proc. SPIE* 772 (1987): 202–9.
9. Progler, C. "An Advanced Method for Lithographic Lens Analysis and Qualification." *Texas Instrum. Tech. J.* 14, no. 3 (1997): 42–50.
10. Seidel, L. *Astr. Nachr.* 43, no. 1027, (1853).
11. Born, M., and E. Wolf. *Principles of Optics*. 6th ed. Oxford, U.K.: Pergamon Press, 1980.
12. Kingslake, R. *Lens Design Fundamentals*. Orlando, FL: Academic Press, 1978.
13. Kingslake, R. *Optical System Design*. Orlando, FL: Academic Press, 1983.
14. Bossung, J. W. "Projection Printing Characterization." *Proc. SPIE* 100 (1977): 80–4.
15. Brunner, T. A. "Impact of Lens Aberrations on Optical Lithography." *IBM J. Res. Dev.* 41, no. 1/2 (1997): 57–67.
16. Flagello, D., and B. Geh. "Lithographic Lens Testing: Analysis of Measured Aerial Images, Interferometric Data and Photoresist Measurements." *Proc. SPIE* 2726 (1996): 788–98.
17. Progler, C., and D. Wheeler. "Optical Lens Specifications from the User's Perspective." *Proc. SPIE* 3334 (1998): 256–68.
18. Lai, K., I. Lalovic, B. Fair, A. Kroyan, C. J. Progler, N. Farrar, D. Ames, and K. Ahmed. "Understanding Chromatic Aberration Impacts on Lithographic Imaging." *J. Microlith. Microfab. Microsyst.* 2, no. 2 (2003): 105–11.
19. Nikolaev, N. I., and A. Erdmann. "Rigorous Simulation of Alignment for Microlithography." *J. Microlith. Microfab. Microsyst.* 2, no. 3 (2003): 220–6.
20. de Zwart, G., M. van den Brink, R. George, D. Satriasaputra, J. Baselmans, H. Butler, J. van Schoot, and J. de Klerk. "Performance of a Step and Scan System for DUV Lithography." *Proc. SPIE* 3051 (1997): 817–35.
21. Magome, N., and H. Kawai. "Total Overlay Analysis for Designing Future Aligner." *Proc. SPIE* 2440 (1995): 902–12.
22. Engelstad, R., E. Lovell, G. Dicks, C. Martin, M. Schlax, W. Semke, A. Liddle, and A. Novembre. "Finite Element Modeling of SCALPEL Masks." *Proc. SPIE* 3676 (1999): 128–39.
23. Mikkelson, A., R. Engelstad, E. Lovell, T. Bloomstein, and M. Mason. "Mechanical Distortions in Optical Reticles." *Proc. SPIE* 3676 (1999): 744–56.
24. Coleman, D., P. Larson, A. Lopata, W. Muth, and A. Starikov. "Accuracy of Overlay Measurements: Tool and Mark Asymmetry Effects." *Proc. SPIE* 1261 (1990): 139–61.
25. Perloff, D. S. "A Four Point Electrical Measurement Technique for Characterizing Mask Superposition Errors on Semiconductor Wafers." *IEEE Solid State Circuits* SC-13, no. 4 (1978): 436–44.

26. Kopp, R. J., and D. J. Stevens. "Overlay Considerations for the Selection of Integrated-Circuit Pattern-Level Sequences." *Solid State Technol.* July (1980): 79–87.
27. Arnold, W., and J. Greeneich. "The Impact of Stepper Overlay on Advanced IC Design Rules." *OCG Microlith. Semin.* (1993): 87–100.
28. Rubingh, R., Y. van Dommelen, S. Tempelaars, M. Boonman, R. Irwin, E. van Donkelaar, H. Burgers., et al. "Performance of a High Productivity 300-mm Dual-Stage 193-nm 0.75-NA TWINSCAN AT:1100B System for 100-nm Applications." *J. Microlith. Microfab. Microsyst.* 2, no. 1 (2003): 8–18.
29. Burnett, J. H., Z. H. Levine, E. L. Shirley, and J. H. Bruning. "Symmetry of Spatial-Dispersion-Induced Birefringence and its Implications for CaF_2 Ultraviolet Optics." *J. Microlith. Microfab. Microsyst.* 1, no. 3 (2002): 213–24.
30. Levenson, M. D. "Wavefront Engineering from 500 to 100 nm CD." *Proc. SPIE* 3051 (1997): 2–13.
31. Fukuda, H., T. Terazawa, and S. Okazaki. "Spatial Filtering for Depth of Focus and Resolution Enhancement in Optical Lithography." *J. Vac. Sci. Technol.* B9, no. 6 (1991): 3113–6.
32. Smith, B. W. "Mutual Optimization of Resolution Enhancement Techniques." *J. Microlith. Microfab. Microsyst.* 1, no. 2 (2002): 95–105.
33. Levenson, M. D., N. S. Viswanathan, and R. A. Simpson. "Improving Resolution in Photolithography with a Phase Shifting Mask." *IEEE Trans. Electron Devices* ED-29, no. 12 (1982): 1812–46.
34. Shibuya, M. Projection master for transmitted illuminated. Japanese Patent Gazette No. 62-50811.
35. Smith, H. I., E. H. Anderson, and M. L. Schattenburg. Lithography mask with a π-phase shifting attenuator. U.S. Patent 4, 890, 309.
36. Flanders, D. C. and H. I. Smith. Spatial period division exposing, U.S. Patent 4, 360, 586.
37. Pierrat, C., A. Wong, and S. Vaidya. "Phase-Shifting Mask Topography Effects on Lithographic Image Quality." *IEDM Tech. Dig.* (1992): 53–6.
38. Tsujimoto, E., T. Watanabe, Y. Sato, A. Moniwa, Y. Igarashi, and K. Nakai. "Hierarchical Mask Data Design System (PROPHET) for Aerial Image Simulation, Automatic Phase-Shifter Placement, and Subpeak Overlap Checking." *Proc. SPIE* 3096 (1997): 163–72.
39. Liu, H.-Y., L. Karklin, Y.-T. Wang, and Y. C. Pati. "The Application of Alternating Phase-Shifting Masks to 140 nm Gate Patterning (I): Linewidth Control Improvements and Design Optimization." *Proc. SPIE* 3236 (1997): 328–37.
40. Liu, H.-Y., L. Karklin, Y.-T. Wang, and Y. C. Pati. "The Application of Alternating Phase-Shifting Masks to 140 nm Gate Patterning (II): Mask Design and Manufacturing Tolerances." *Proc. SPIE* 3334 (1998): 2–14.
41. Terazawa, T., N. Hasegawa, H. Fukuda, and S. Katagiri. "Imaging Characteristics of Multi-Phase-Shifting and Halftone Phase-Shifting Masks." *Jpn. J. Appl. Phys.* 30 (1991): 2991–7.
42. Ma, Z. M., and A. Andersson. "Preventing Sidelobe Printing in Applying Attenuated Phase-Shift Reticles." *Proc. SPIE* 3334 (1998): 543–52.
43. Cuthbert, J. D. "Optical Projection Printing." *Solid State Technol.* August (1977): 59–69.
44. Herschel, R. S. "Partial Coherence in Projection Printing." *Proc. SPIE* 135 (1987): 24–9.
45. Smith, B. W. "Revalidation of the Rayleigh Resolution and DOF Limits." *Proc. SPIE* 3334 (1998): 142–53.
46. Smith, B. W. "Optics for Photolithography." In *Microlithography: Science and Technology*, edited by J. R. Sheats, and B. W. Smith, New York: Marcel Dekker, 1998, chap. 3.
47. Shiraishi, N., S. Hirukawa, Y. Takeuchi, and N. Magome. "SHRINC: A New Imaging Technique for 64 Mbit DRAM." *Microlith. World* July (1992): 7–14.
48. Shiraishi, N., S. Hirukawa, Y. Takeuchi, and N. Magome. "New Imaging Technique for 64 M-DRAM." *Proc. SPIE* 1674 (1992): 741–52.
49. Granik, Y. "Source Optimization for Image Fidelity and Throughput." *J. Microlith. Microfab. Microsyst.* 3, no. 4 (2004): 509–22.
50. Rogoff, R., G. Davies, J. Mulkens, J. de Klerk, P. van Oorschot, G. Kalmbach, J. Wangler, and W. Rupp. "Photolithography Using the AERIAL Illuminator in a Variable-NA Wafer Stepper." *Proc. SPIE* 2726 (1996): 54–70.
51. Hopkins, H. H. *Proc. R. Soc.* A208 (1951): 263.

52. Pforr, R., A. Wong, K. Ronse, L. Van den hove, A. Yen, S. Palmer, G. Fuller, and O. Otto. "Feature Biasing versus Feature-Assisted Lithography—A Comparison of Proximity Correction Methods for $0.5\times(\lambda/\mathrm{NA})$ Lithography." *Proc. SPIE* 2440 (1995): 150–70.
53. Kawamura, E., T. Haruki, Y. Manabe, and I. Hanyu. "Simple Method of Correcting Optical Proximity Effect for 0.35 μm Logic LSI Circuits." *Jpn. J. Appl. Phys.* 34 (1995): 6547–51.
54. Bruning, J. H. "Optical Lithography below 100 nm." *Solid State Technol.* November (1998): 59–67.
55. Sturtevant, J. unpublished.
56. Oldham, W. G., W. Arden, H. Binder, and C. Ting. "Contrast Studies in High-Performance Projection Optics." *IEEE Trans. Electron Devices* ED-30, no. 11 (1983): 1474–9.
57. Mack, C. "Understanding Focus Effects in Submicrometer Optical Lithography." *Opt. Eng.* 27, no. 12 (1988): 1093–100.
58. Mack, C. A. "Mask Linearity and the Mask Error Enhancement Factor." *Microlith. World* 8, no. 1 (1999): 11–2.
59. Maurer, W., K. Satoh, D. Samuels, and T. Fischer. "Pattern Transfer at $k_1 = 0.5$: Get 0.25 μm Lithography Ready for Manufacturing." *Proc. SPIE* 2726 (1996): 113–24.
60. Wong, K., R. A. Ferguson, L. W. Liebmann, S. M. Mansfield, A. F. Molless, and M. O. Neiser. "Lithographic Effects of Mask Critical Dimension Error." *Proc. SPIE* 3334 (1998): 106–16.
61. Schellenberg, F. M., V. Boksha, N. Cobb, J. C. Lai, C. H. Chen, and C. Mack. "Impact of Mask Error Factors on Full Chip Error Budgets." *Proc. SPIE* 3679 (1999): 261–75.
62. van Schoot, J., J. Finders, K. van Ingen Schenau, M. Klaassen, and C. Buijk. "The Mask Error Factor: Causes and Implications for Process Latitude." *Proc. SPIE* 3679 (1999): 250–60.
63. Granik, Y. "Generalized Mask Error Enhancement Factor Theory." *J. Microlith. Microfab. Microsyst.* 4, no. 2 (2005): 023001.
64. Mason, M. private communication.
65. Maltabes, J., M. Hakey, and A. Levine. "Cost/Benefit Analysis of Mix-and-Match Lithography for Production of Half-Micron Devices." *Proc. SPIE* 1927 (1993): 814–26.
66. Burggraaf, P. "Applying Cost Modeling to Stepper Lithography." *Semicond. Int.* February (1994): 40–4.
67. Wong, A. K., A. F. Molless, T. A. Brunner, E. Coker, R. H. Fair, G. L. Mack, and S. M. Mansfield. "Linewidth Variation Characterization by Spatial Decomposition." *J. Microlith. Microfab. Microsyst.* 1, no. 2 (2002): 106–16.
68. Owa, S., and H. Nagasaka. "Immersion Lithography; Its Potential Performance and Issues." *Proc. SPIE* 5040 (2003): 724–33.
69. Owa, S., H. Nagasaka, Y. Ishii, O. Hirakawa, and T. Yamamoto. "Feasibility of Immersion Lithography." *Proc. SPIE* 5377 (2004): 264–72.
70. Owa, S., H. Nagasaka, Y. Ishii, K. Shiraishi, and S. Hirukawa. "Fullfield Exposure Tools for Immersion Lithography." *Proc. SPIE* 5754 (2005): 655–68.
71. Owa, S., H. Nagasaka, K. Nakano, and Y. Ohmura. "Current Status and Future Prospect of Immersion Lithography." *Proc. SPIE* 6154 (2006): 615408-1–615408-12.
72. Okumura, M., J. Ishikawa, M. Hamatani, and M. Nei. "Mass Production Level ArF Immersion Exposure Tool." *Proc. SPIE* 6154 (2006): 61541U-1–61541U-8.
73. Jasper, H. C., T. Modderman, M. van de Kerkhof, C. Wagner, J. Mulkens, W. de Boeij, E. van Setten, and B. Kneer. "Immersion Lithography with an Ultrahigh-NA In-Line Catadioptric Lens and a High-Transmission Flexible Polarization Illumination System." *Proc. SPIE* 6154 (2006): 61541W-1–61541W-14.
74. Lin, B. J. "Immersion Lithography and Its Impact on Semiconductor Manufacturing." *J. Microlith. Microfab. Microsyst.* 3, no. 3 (2004): 377–95.
75. Lin, B. J. "Semiconductor Foundry, Lithography, and Partners." *Proc. SPIE* 4688 (2002): 11–24.
76. Owa, S., and H. Nagasaka. "Advantage and Feasibility of Immersion Lithography." *J. Microlith. Microfab. Microsyst.* 3, no. 1 (2004): 97–103.
77. Mulkens, J., D. Flagello, B. Streefkerk, and P. Graeupner. "Benefits and Limitations of Immersion Lithography." *J. Microlith. Microfab. Microsyst.* 3, no. 1 (2004): 104–14.

78. Ohmura, Y., M. Nakagawa, T. Matsuyama, and Y. Shibazaki. "Catadioptric Lens Development for DUV and VUV Projection Optics." *Proc. SPIE* 5040 (2003): 781–8.
79. Ikezawa, H., Y. Ohmura, T. Matsuyama, Y. Uehara, and T. Ishiyama. "A Hyper-NA Projection Lens for ArF Immersion Exposure Tool." *Proc. SPIE* 6154 (2006): 615421-1–615421-8.
80. Smith, B. W., A. Bourov, H. Kang, F. Cropanese, Y. Fan, N. Lafferty, and L. Zavyalova. "Water Immersion Optical Lithography at 193 nm." *J. Microlith. Microfab. Microsyst.* 3, no. 1 (2004): 44–51.
81. Burnett, J. H., and S. Kaplan. "Measurements of the Refractive Index and the Thermo-Optic Coefficient for Water around 193 nm." *Proc. SPIE* 5040 (2003): 1742–9.
82. Flagello, D., B. Geh, S. Hansen, and M. Totzeck. "Polarization Effects Associated with Hyper-Numerical-Aperture (>1) Lithography." *J. Microlith. Microfab. Microsyst.* 4, no. 3 (2005): 031104.
83. Adam, K., and W. Maurer. "Polarization Effects in Immersion Lithography." *J. Microlith. Microfab. Microsyst.* 4, no. 3 (2005): 031106.
84. Matsuyama, T., and T. Nakashima. "Study of High NA Imaging with Polarized Illumination." *Proc. SPIE* 5754 (2005): 1078–89.
85. Brunner, T. A., N. Seong, W. D. Hinsburg, J. A. Hoffnagle, F. A. Houle, and M. I. Sanchez. "High Numerical Aperture Lithographic Imagery at the Brewster Angle." *J. Microlith. Microfab. Microsyst.* 1, no. 3 (2002): 188–96.
86. Lin, B. J. "Simulation of Optical Projection with Polarization-Dependent Stray Light to Explore the Difference between Dry and Immersion Lithography." *J. Microlith. Microfab. Microsyst.* 3, no. 1 (2004): 9–20.
87. Estroff, A., Y. Fan, A. Bourov, and B. Smith. "Mask-Induced Polarization Effects at High Numerical Aperture." *J. Microlith. Microfab. Microsyst.* 4, no. 3 (2005): 031107.
88. Totzeck, M., P. Gräupner, T. Heil, A. Göhnermeier, O. Dittmann, D. Krähmer, V. Kamenov, J. Ruoff, and D. Flagello. "Polarization Influence on Imaging." *J. Microlith. Microfab. Microsyst.* 4, no. 3 (2005): 031108.
89. Moore, G. E. "Lithography and the Future of Moore's Law." *Proc. SPIE* 2440 (1995): 2–17.
90. Mack, C. "Trends in Optical Lithography." *Opt. Photo. News.* April (1996): 29–33.
91. Brueck, S., and X. Chen. "Spatial Frequency Analysis of Optical Lithography Resolution Enhancement Techniques." *Proc. SPIE* 3679 (1999): 715–25.

19

Photoresist Materials and Processing

19.1 Formation of the Relief Image **19**-1
Overview • Description of the Lithographic Process

19.2 Formation of a Relief Image in Novolac-Based Photoresists **19**-5
Overview • Elements of the Dissolution Mechanism of Novolac-Based Photoresists • Development Mechanisms in Novolac-Based Photoresists

19.3 Formation of the Relief Image in Chemically Amplified Resists..... **19**-23
Overview • Exposure Step

19.4 ArF Materials, Immersion Lithography and Extension of ArF..... **19**-40
ArF Materials • ArF Transparent Polymer Systems • Extending ArF • Topcoats for Immersion Lithography • New Immersion Fluids • High Refractive Index (RI) Polymers • Post-ArF-Material Requirements

References **19**-53

César M. Garza
Will Conley
Freescale Semiconductor, Inc.

Jeff Byers
KLA-Tencor

19.1 Formation of the Relief Image

Optical microlithography is the technology that determines, in practical terms, the smallest transistor dimensions that can be manufactured on a semiconductor chip. As such it has been the primary driver for the remarkable improvements in performance and reduction in cost per function, the hallmark of the microelectronics industry. Optical microlithography involves the practice of multiple disciplines: physics, chemistry, and engineering specialties. Physics is used to form the aerial image; and it has been covered in the previous chapter. Chemistry is involved in the formation of the latent and relief images on the recording medium, know as photoresist, and it is the subject matter of the present chapter.

19.1.1 Overview

As it was covered in the previous chapter, the smallest dimension that be printed is given by the Rayleigh criteria:

$$\text{Resolution} = k_1 \lambda / \text{NA} \tag{19.1}$$

where λ is the actinic wavelength used in the formation of the aerial image, k_1 is a proportionality

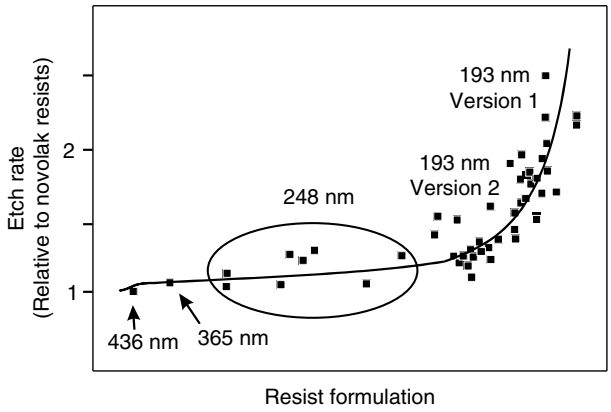


FIGURE 19.1 Etch rates of 365, 248, and 193 nm resists relative to 436 nm resists.

constant, and NA is the numerical aperture of the lens. The proportionality constant, k_1, can be used to assess the maturity of the process; the theoretical limit is 0.25.

The path that the industry has followed to improve resolution has been to first increase the NA of the lens to its practical limit; and then reduce the wavelength. This has deep implications in the formation of the relief image, for the resist chemistry is optimized for a specific wavelength. The wavelengths that have been used in optical microlithography are: (a) 436 nm, which corresponds to the g-line of a mercury lamp; (b) 365 nm, which corresponds to the i-line of a mercury lamp; (c) 248 nm, which corresponds to a KrF excimer laser; and (d) 193 nm, which corresponds to an ArF excimer laser. Photoresists used on the first two wavelengths, 436 and 365 nm, are made using the same basic chemistry, and it involves using a novolak-resin and a diazoquinone sensitizer. Despite the similarity in their chemistry, the resist formulations at 436 and 365 nm are different because they need to be optimized for each wavelength. This chemistry is covered in Section 19.2.

Because of low intensity at the resist level, a completely new technology called chemical amplification had to be developed to formulate the resists at 248 and 193 nm. Like in the case of 436 and 365 nm, the resist formulations at 248 and 193 nm are different because they need to be optimized for a specific wavelength. The chemistry for these resists is covered in Section 19.3 and Section 19.4. One unfortunate drawback in moving from a novolak-resist to a chemically amplified resist (CAR) formulation is a decrease[1] in etch resistance, one of the primary qualities of interest in a photoresist. This is shown in Figure 19.1, where we plot the etch rate of 248 and 193 nm relative to that of novolak-based photoresists. Notwithstanding this drawback, and others that will be mentioned later in the chapter, the performance of resists in terms of resolution has improved over time.[2]

This point is made in Figure 19.2, where we show a plot of the proportionality constant, k_1, over time. This shows that the improvement in resolution-realized overtime has been greater than what can be expected by increasing the NA or decreasing the wavelength. A number of factors have contributed to this improvement, like better equipment and control methodologies. But the resist formulations have also improved, because acceptable resist patterns are being printed despite a decrease in contrast on the aerial image.

19.1.2 Description of the Lithographic Process

A schematic representation of the microlithographic process is shown in Figure 19.3, and a flowchart is shown in Figure 19.4. Excluding the hard bake, which is not always required, the formation of the relief image is the last step and goal of the microlithographic process. The relief image is formed by developing

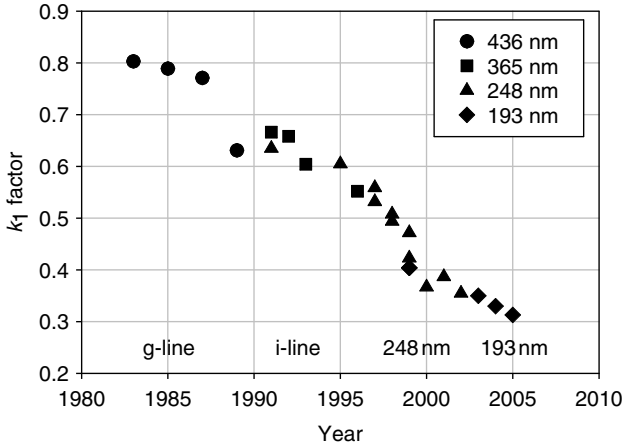


FIGURE 19.2 Proportionality constant, k_1, of the Rayleigh criterion (Equation 19.1) plotted vs. time, showing an improvement of the lithographic process over time.

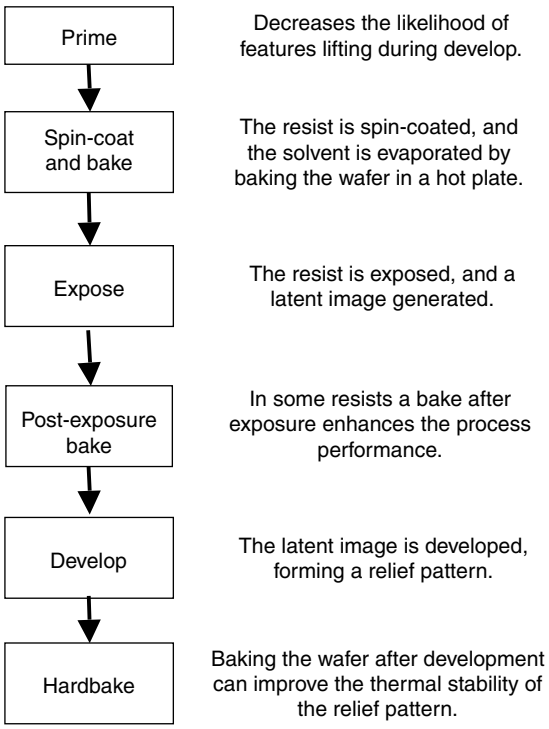

Prime
Decreases the likelihood of features lifting during develop.
Spin-coat and bake
The resist is spin-coated, and the solvent is evaporated by baking the wafer in a hot plate.
Expose
The resist is exposed, and a latent image generated.
Post-exposure bake
In some resists a bake after exposure enhances the process performance.
Develop
The latent image is developed, forming a relief pattern.
Hardbake
Baking the wafer after development can improve the thermal stability of the relief pattern.

FIGURE 19.3 Flowchart of the microlithographic process. The postexposure and hard-bake steps can be omitted, depending on the process.

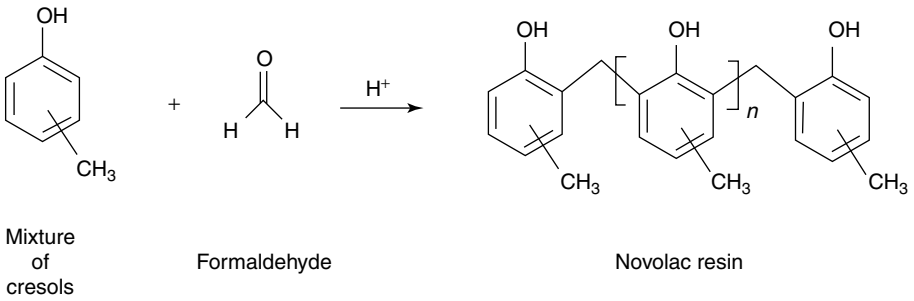


FIGURE 19.4 Schematic representation of the formation of novolac resins.

the latent image, which is generated during the exposure step in an appropriate medium. This medium is called the developer. The developer is usually a liquid, but it can also be a plasma gas.

Over the years, a large number of chemistries have been proposed for optical photoresists, but only a few of them have gained wide acceptance in their practical use. Regardless of how different the chemistries might be, all of them have these two common characteristics:

1. A base material, usually an organic resin, provides the required physical and mechanical properties. In broad terms, these properties are the ability to form very uniform films and good thermal, mechanical, and etch stability.
2. Upon exposure, at least one of the components in the photoresist must undergo one or more chemical changes. The end result is that in the appropriate developer, the exposed-resist areas turn either more soluble (positive-tone) or less soluble (negative-tone) than the unexposed-resist areas. The change in solubility induced by the photochemical change is what permits the formation of the relief image. This principle is responsible for the formation of relief images in all photoresists; however, given the degree of sophistication, the current, and the demands on the photoresist performance, many variables must be carefully controlled for the successful generation of relief images.

Photoresist systems have been classified as positive-tone or negative-tone. Positive-tone photoresist systems are those, where the relief image formed at the wafer level is the same as the one in the mask or reticle. In negative-tone systems, the relief image is the complement or opposite of the mask or reticle. Resists can also be classified as one- or two-component, according to the number of active materials. For instance, if the same material that is sensitive to light also provides the required physical, chemical, and mechanical properties, the photoresist is said to be a one-component system. If two materials are needed, it is said to be a two-component system, and so on. Traditionally, photoresist systems have also been classified according to the form of development. A wet-developed system is one where a liquid is used as the developer, and a dry-developed system is one where the relief image is formed in a plasma gas. By far, wet-development is the most common means of forming the relief image.

The specific mechanism by which the relief image is formed has significant implications on the processing characteristics of the photoresist. The very first photoresists used in the microelectronics industry, before novolak-based photoresists, were rubber-based. In this kind of material, a cross-linking reaction that takes place in the exposed areas increases the molecular weight. The increased molecular weight, in turn, decreases the solubility of the exposed areas in a non-polar, organic solvent such as toluene or xylene. The solubility differential permits the formation of the relief image. On this kind of chemistry, lifting of resist features during development is not an issue nor is the physical and chemical stability of the printed images during the subsequent steps, particularly wet-etching. The reason is that the in situ cross-linking reaction produces a very stable polymer with very good adhesion to the substrate.

However, it also has the detrimental side effect of trapping solvent and monomer molecules inside the cross-linked structure. This leads to a swelling effect that makes this type of processing uncontrollable for printing features smaller than 1.5 μm. Reducing the minimum features and controlling their width became an intractable problem that eventually made these resists obsolete in state-of-the-art wafer fabs.

Positive-tone, novolac-based photoresists with a diazonaphthoquinone (DNQ) sensitizer began replacing rubber-based photoresists in the mid 1970s as greater resolution was needed. A large change in the dissolution rate of the exposed vs. the unexposed resist is the fundamental principle that makes all novolac photoresists work. Very high resolution is possible, almost all the way to the molecular level, but these resists do not have the same adhesion properties of cross-linked materials. The wafers now need to be primed to address this problem. The etch resistance is somewhat inferior but still sufficient, particularly, in a plasma etch. The same basic chemistry works at 436 and 365 nm, but the formulation is different, mainly in the sensitizer and resin composition, to address the higher absorption of the 436-nm materials at 365 nm.

Because the trend for most organic materials is to become more absorptive as the wavelength decreases, new materials had to be developed with the advent of 248-nm lithography. Another significant problem that had to be overcome is a much lower number of photons available at 248 nm as compared to 365 nm. This led to the development of new families of materials generically known as chemically amplified photoresists. In those type of materials, like novolac-based photoresists, the dissolution rate of the exposed is much higher than the unexposed resist. The way this is accomplished, however, is completely different. In a chemical amplified resist, there is a chemical deprotection reaction on the resin that is catalytic in nature. In this deprotection reaction, a segment of the original resin is removed as a gas. This can lead to a contamination-deposition on the imaging tool, if it is not properly managed. The deprotection reaction can be affected by airborne contaminants, and it can also be induced by other means, like temperature and exposure to an electron beam. This requires a very tight control of the environment, temperature, and timing between exposure and development. The way to address these problems was to couple the coater-developer unit with the exposure tool, so that the two works as a single unit, known as a cell. A further decrease in the adhesion and etch resistance of chemically amplified resist can also be traced by these deprotection reaction.

Absorption again became a problem when the wavelength had to be reduced from 248 down to 193 nm to further reduce the transistor size. Chemical-amplification is also used in the formulation 193 nm, but the chemical composition of the resin and sensitizer is different. The etch resistance of 193 nm is lower yet than that of 248 nm. The new issue with 193 nm is line edge roughness (LER). The printed resist images are rougher compared to previous materials, although this problem is being resolved by further refining the resist formulations. The resolution of 193-nm lithography is likely to be extended to the 45-nm node by the used of a liquid to bend the light rays, so that in practice, the NA is greater than one. This undoubtedly will require further changes to the resist formulation.

19.2 Formation of a Relief Image in Novolac-Based Photoresists

19.2.1 Overview

Novolac-based photoresists have been classified as two-component photoresist systems[3,4] because the two main components are a novolac resin and DNQ sensitizer. In more advanced photoresist formulations, this can be an oversimplification due to the presence of other important additives such as dyes to minimize reflections from the substrate and surfactants to improve the coating uniformity. Nevertheless, it still makes sense to keep this classification because the resin and the sensitizer play the most important roles in the formation of the relief image.

Novolac resins are phenolic resins formed by the condensation of various cresols with formaldehyde, as shown in Figure 19.5. They are very suitable as a basis for the formulation of photoresists because they can form very uniform, thin films when spun-coat. Also, the aromatic rings that form the resin backbone produce a great deal of chemical stability, allowing the resist to withstand the harsh environments

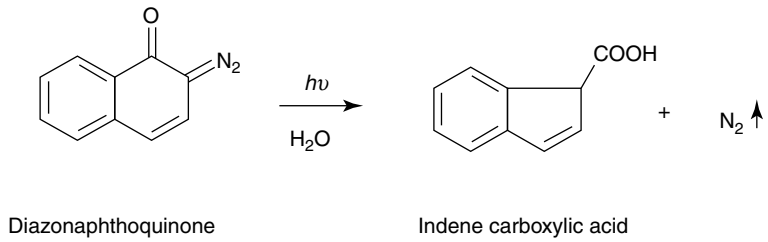


FIGURE 19.5 Photochemical reaction of a diazonaphthoquinone (DNQ).

encountered in subsequent processing steps. This is particularly important when the next step is a pattern transfer step like plasma etching. Even though these are important contributions to the overall requirements of a photoresist system, the role of the resin goes beyond that; the resin structure has a great deal of influence in the dissolution process. As such, it has a large effect on the formation of the relief image.

Since novolac resins are not sensitive to light, a second component is needed. This is the role of the sensitizer, which is a DNQ sulfonate derivative. Upon exposure, it undergoes a photochemical transformation, the end result of which is the formation of an indene-carboxylic acid, or photoacid (see Figure 19.6).

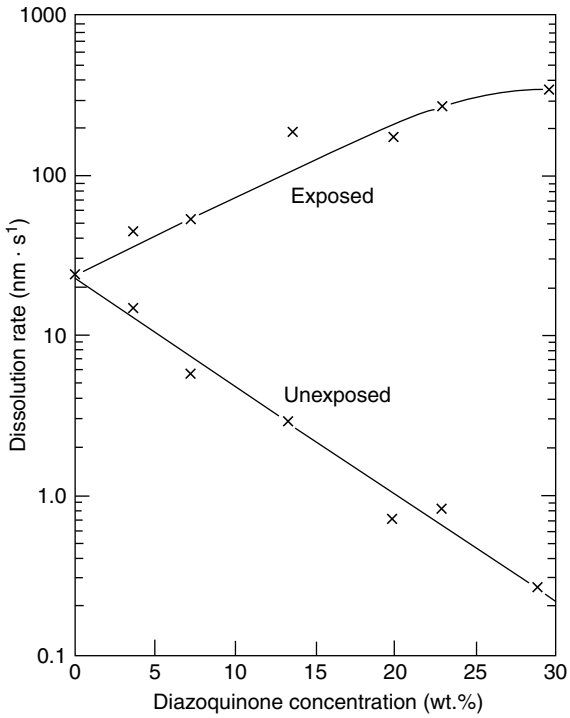


FIGURE 19.6 Dissolution rates of exposed and unexposed photoresists. (From *Introduction to Microlithography*, edited by Thompson, L. F., Willson, C. G., and Bowden, M. J., ACS Symposium Series 219, American Chemical Society, Washington, DC, 90–91, 1983. Copyright IEEE, 1980.)

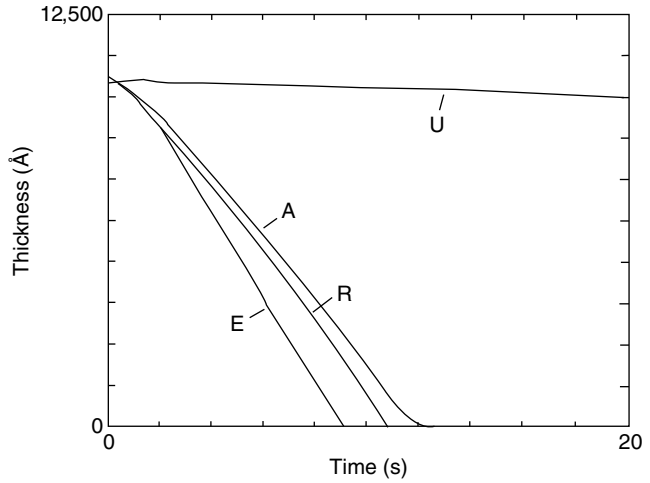


FIGURE 19.7 Dissolution rates of exposed, E, and unexposed, U, photoresist; novolac resin, R; and novolac resin mixed with an indene-carboxylic acid, A. (From *Introduction to Microlithography*, edited by Thompson, L. F., Willson, C. G., and Bowden, M. J., ACS Symposium Series 219, American Chemical Society, Washington, DC, 111–16, 1983.)

Novolac resins are soluble in aqueous alkali, but the addition of a DNQ sulfonate inhibits this dissolution process. Figure 19.7[5] shows, in a graphic way, the dramatic effect of adding a DNQ sulfonate on the dissolution rates of novolac resins. From this plot, we can see that the difference in the dissolution rates between exposed and unexposed photoresist can be as much as a factor of 200.

The photoacid, on the other hand, not only does not inhibit the dissolution process, but it may also actually enhance it, although this is somewhat controversial because of contradictory reports in the literature. Hinsberg et al.[6] measured the dissolution rates of fast-dissolving resins and fast-dissolving resins mixed with photoacids. A summary of their data is shown in Figure 19.8. On this

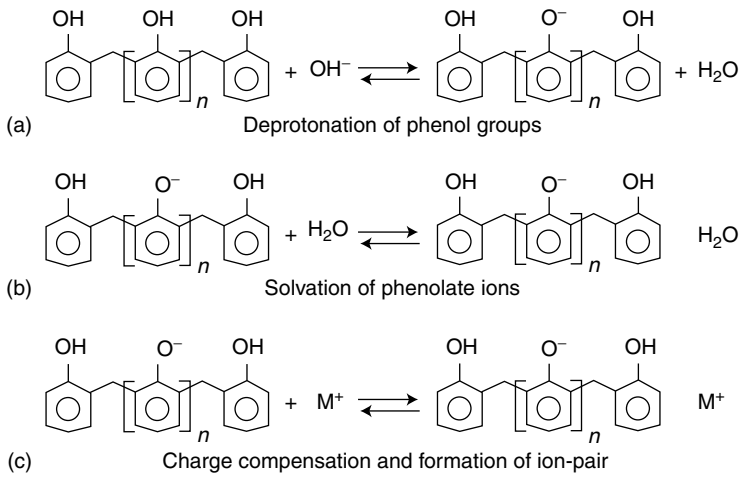


FIGURE 19.8 (a) Deprotonation of phenol groups, (b) Solvation of phenolate ions, (c) Charge compensation and formation of ion-pair.)

data, we see that the exposed photoresist dissolves faster than the original resin, although the resin mixed separately with the acid does not. It could be as Ouano[7] proposes that the evolution of nitrogen gas, a byproduct of the photochemical reaction shown in Figure 19.9, creates free volume that facilitates the diffusion of the developer into the resin matrix, enhancing the dissolution process. On the other hand, Blum et al.[8] report a significant increase of the dissolution rates for slower-dissolving resins mixed with a photoacid.

Whether the presence of the photoacid enhances the dissolution rate of the novolac resin or not, is not as important as identifying the two determining factors for the formation of the relief image:

1. The difference in the dissolution rates between the exposed and unexposed photoresists.
2. The rate of change of the dissolution rates as a function of change of the amount of irradiated light.

The nature of these two factors is clearly kinetic and provides the foundation upon which the formation of the relief image rests. It is of vital importance to determine the factors that has an effect on these two phenomena, not only to develop optimum resist formulations, but also to optimize and control the process in a manufacturing environment. For many years, the industry approached this goal in a highly empirical manner. Fortunately in recent years, great progress has been made in understanding the mechanism of the dissolution process and modeling of the overall lithographic process.

The mechanism of a chemical reaction is the detailed description of the step or steps that lead to the formation of the chemical products from the starting materials. In our case, the starting materials are the resin, the sensitizer (the two major photoresist components), light, and the developer. The product is the solvated, ionized resin, i.e., the dissolution of the novolac resin in the developer. The benefits of having a clear and detailed understanding of the mechanism of the dissolution process should be obvious.

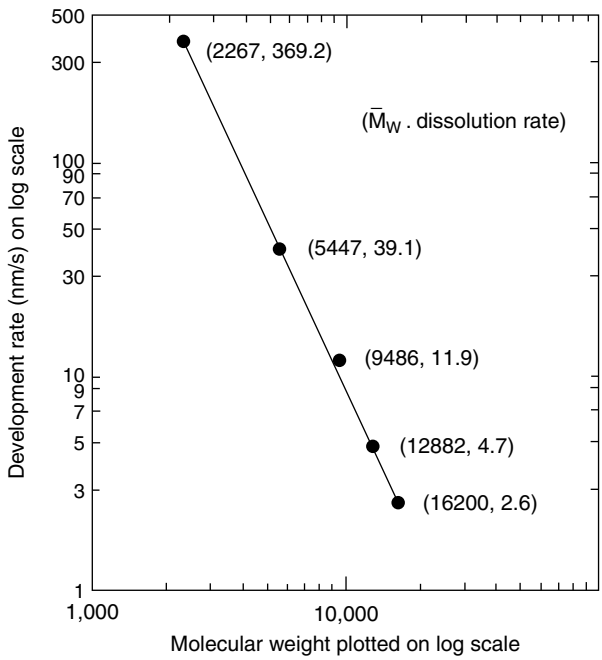


FIGURE 19.9 Effect of the molecular weight on the dissolution rate of photoresists. (From Ouano, A. C., *Polym. Eng. Sci.*, 18, 306, 1978.)

To the resist manufacturer, it provides a solid base from which to manipulate the chemical quality and quantity of the individual components required to optimize the resist formulation. To the process engineer, it provides the foundation for selecting the best processing conditions and troubleshooting.

19.2.2 Elements of the Dissolution Mechanism of Novolac-Based Photoresists

As previously stated, the mechanism of a chemical reaction is the detailed description of the steps involved in the chemical transformation of interest. In our case, the transformation of interest is the dissolution of the novolac resin into the aqueous developer. The generally accepted steps involved in the dissolution process are:

1. Diffusion of –OH ions and water into the polymer matrix.
2. Deprotonation of phenol groups of the novolac resin to polymer-bound phenolate ions, as shown in Figure 19.10a.
3. Solvation of the phenolate ions, as shown in Figure 19.10b.
4. Compensation of the negative charge of the phenolate ions by the positive charge of the base cation and formation of ion pairs, as shown in Figure 19.10c.
5. Rearrangement of the ionized polymer chains, detachment from the polymer matrix, and transfer into solution.

The structures of the resin, the sensitizer, and the developer are all involved in this multi-step process. It is convenient to start by analyzing the contributions of each of these variables before examining the proposed dissolution mechanisms.

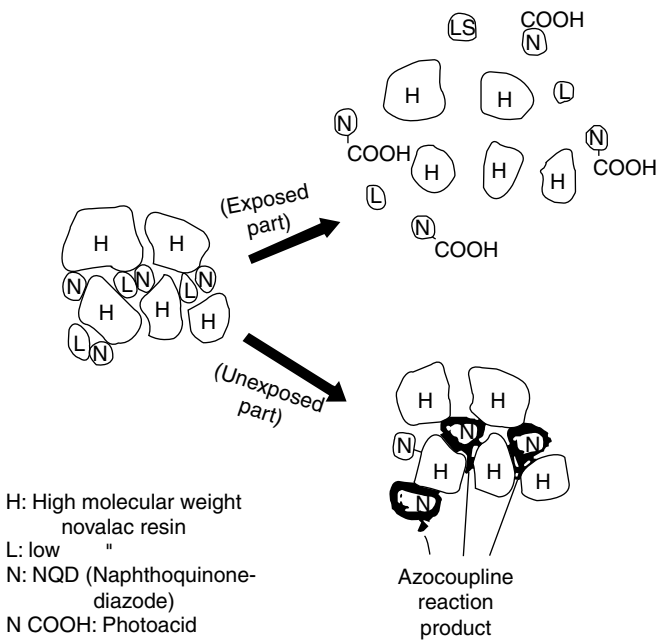


FIGURE 19.10 Schematic representation of the stone-wall model. (From Blum, L., Perkins, M. E., and McCullough, A. W., *Proc. SPIE*, 771, 148, 1987.)

19.2.2.1 Novolac Resin

The molecular weight, Mw, of the novolac resins used in photoresists usually lies in the range 1000–3000, corresponding to 8–25 repeating units, *n*, in Figure 19.4. As a general rule, the higher the molecular weight of the resin, the slower is the dissolution rate of the photoresist. This trend is clearly shown in Figure 19.9, where Arcus[9] plots the dissolution rates of a group of photoresists that vary only by their average molecular weight. The exceptions to this general rule mainly relate to the isomeric composition of the resin, as we shall see later in this section.

The slower dissolution rate of higher-molecular-weight resins tends to degrade the photoresist performance in particular resolution.[10,11] However, high-molecular-weight components can enhance other required photoresist properties, like thermal stability, because their glass transition temperature is higher. Since Mw has an effect on the photoresist performance, intuitively one would also think that its variation, or molecular dispersity, Mw/Mn, would also have an effect on the photoresist performance. This is shown in Table 19.1, where the contrast and photosensitivity are tabulated for photoresists made from resins with different Mw and Mw/Mn. If the photoresist performance is affected by Mw and Mw/Mn, the repeatability of the process over time clearly would depend on how well Mw and Mw/Mn are controlled. This is an important point that needs to be made: microlithographic engineers place as much importance on the repeatability as in performance of the process itself.

A different approach to narrowly controlling the molecular-weight dispersity to enhance the performance of novolac-based photoresists is advocated by Hanabata et al.[12] This approach is best known as stone-wall model, which invokes the formation of a base-catalyzed azo-coupling between the DNQ and the phenolic resin in the unexposed areas of the resist film. The model derives its name from the analogy made between the photoresist structure and a "stone wall". The azo-coupling reaction causes an increase in molecular weight and, thereby, creates a stone wall that is resistant to dissolution in base. The lower-molecular-weight components are like the small stones that fit in-between the larger stones, or higher-molecular-weight components, in a stone-wall-like structure. After the photoresist has been exposed and the DNQ turned into an acid, the lower-molecular-weight components will readily dissolve. This will bring about the collapse of the wall, which in turn will facilitate the dissolution of the higher-molecular-weight components, since they will be surrounded by the developer. Figure 19.10 graphically describes this model.

The stone-wall model is intuitively appealing and complements nicely some more recent models, such as the domain theory. One advantage of this approach is blending: the physical properties of the different-molecular-weight components, such as dissolution rate and thermal stability can be blended together, producing a photoresist with the best compromise in terms of resolution and temperature stability. This model describes many aspects of photoresist–dissolution response, but it has fallen from favor because many compounds that have been demonstrated to have a powerful dissolution–inhibition response do not undergo azo-coupling. The 1,3-diacyl-2-diazo compounds described by Grant et al.[13] are functional examples, and the dissolution inhibition of the naphthalene sulfonate of hydroxybenzophenone is essentially equal to that of the corresponding DNQ.

TABLE 19.1 Photosensitivity and Contrast of Photoresists Made from Resins with Different Average Molecular

Novolac Resin	Weight (g)	$\overline{M}w$	$\overline{M}/\overline{M}n$	Resist Visc. (cSt)	Devel. Strength	Erosion Rate (Å/min)	Photosensitivity (mJ/cm^2)	Contrast
Unfractionated	350	19,100	72.1	32.3	1:1	10	109	1.66
C	104	45,800	70.0	120.7	Conc.	47	167	1.71
D	59	12,300	38.7	21.1	1:1	16	149	1.65
E	101	4,870	20.8	11.0	1:4	32	142	1:10
F	38	540	6.8		Not made into a resist			

Weight, $\overline{M}w$, and dispersity, $\overline{M}w/\overline{M}n$.

Source: Reproduced from Pampalone, T. R., *Solid State Technol.*, 27(6), 115, 1984. With permission.

The isomeric composition of the resin also plays a very important role in the photoresist performance. Isomers are chemicals with the same formula, but with different configurations in space. Cresol, the starting material for novolac resins, has three isomers: *ortho*-, *meta*-, and *para*-cresol (*o*-cresol, *m*-cresol and *p*-cresol, for short). They are shown in Figure 19.11. Notice that the only difference between the three is the relative position of the –OH and $–CH_3$ groups in the phenyl ring.

Since, cresol is the starting material for the manufacture of novolac resins, the relative position of the methylene links within the resin will vary, depending on which cresol isomer is involved in the polymerization reaction. The implication of this phenomenon is that the three-dimensional, or secondary structure of the resin will greatly vary for each one of these isomeric novolac resins. This was the subject of a landmark paper by Templeton, Szmanda, and Zampini.[14] According to this study, *ortho–ortho* linked phenolic polymers (e.g., *p*-cresol novolac) demonstrate considerable intramolecular hydrogen bonding, whereas (pHOST) and *ortho–para* linked novolac polymers display primarily intermolecular hydrogen bonding. Pawloski et al.[15] reached the same conclusion regarding secondary structure in their recent molecular-dynamics study of the clustering of hydroxyl groups in phenolic polymers. The two- and three-dimensional structure of an *ortho–ortho* novolac trimer is shown in Figure 19.12. Note that the two-dimensional representation of the trimer does not provide information on the close proximity of the –OH groups to each other. This close proximity leads to strong intermolecular hydrogen bonding. On the other hand, an *ortho–para*-coupled trimer such as the one shown in Figure 19.13, has a three-dimensional structure, where the –OH groups are directed outward. In this case intramolecular, and not intermolecular, hydrogen bonding takes place. Templeton et al. showed that the bulk dissolution rate of resins, where intermolecular hydrogen bonding is prevalent is significantly higher than the dissolution rate of resins, where intramolecular hydrogen bonding prevails.

The secondary structure of *ortho–ortho* novolac resins allows stronger interactions with the sensitizer via hydrogen bonds than those that appear with *ortho–para* novolac resins. The reason is the location in space of the –OH groups and their close proximity to each other. Experimental evidence of this interaction has been reported in the literature[16] as early as 1988, by pointing out the shift towards the red in the infrared spectrum of DNQs after mixing them with novolac resin. However, it was Honda[17] who clearly proposed this interaction in his domain theory, a schematic representation, of which is shown in Figure 19.14. The evidence that Honda et al. provide in support of this theory is two-fold: (a) infrared and C-13 NMR spectroscopy, and (b) correlation between the structure of the resin and lithographic performance. A similar model to the domain theory is the Host Guest Complex, postulated independently by Kajita et al.[18] Like Honda, they offer spectroscopic data and lithographic performance as evidence to postulate the existence of strong resin–sensitizer hydrogen bonding as required for optimum lithographic performance.

Borzo et al.[19] presented evidence from NMR and spectroscopic techniques that the spectroscopic shifts reported by Honda et al. reflected a larger Fermi-resonance effect rather than hydrogen bonding between the DNQ and the phenolic resins. Shifts in the 15 N spectra were small compared to the broadening of the resonance, and the *ortho–ortho* bonded resins did not show differences that would exist for

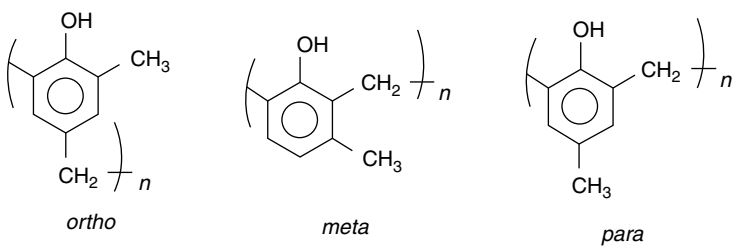
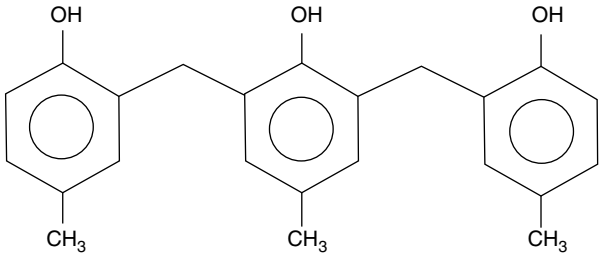
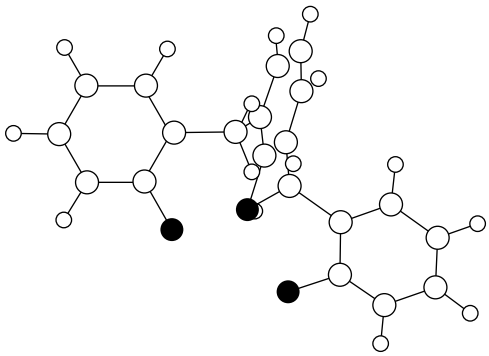


FIGURE 19.11 The *ortho*-, *meta*-, and *para*-cresol isomers.



The two-dimensional structure of an *ortho*, *ortho*' -coupled novolac trimer.



The three-dimensional structure of an *ortho*, *ortho*' -coupled novolac trimer.

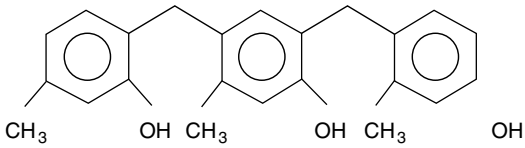
FIGURE 19.12 Comparison between the two- and three-dimensional structure of an *ortho*, *ortho*'-coupled novolac trimer. Note the close proximity between the OH groups that leads to strong intermolecular hydrogen bonding.

macromolecular complexes of the sort described by the Honda's model. These results suggest that the domain theory cannot alone account for the dissolution mechanism.

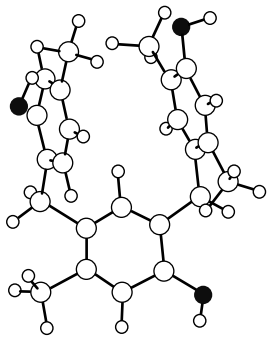
19.2.2.2 Sensitizer

As previously described, the sensitizer a DNQ derivative, inhibits the resin's dissolution process. However, after being exposed to the appropriate wavelength, it turns into an acid and ceases to inhibit the dissolution process. Figure 19.6 shows how remarkable the inhibition effect of the sensitizer is; addition of some 20% in weight of a DNQ derivative slows the dissolution process by more than two orders of magnitude. Clearly, this effect is out of proportion to its concentration, and it implies that the role the inhibitor plays must take place at a critical stage in the development process.

For many years, there has been little information in the literature on the exact sensitizer composition of commercial photoresists, and even less on how they inhibit the dissolution process. This is not very surprising since this kind of information is regarded as a trade secret by resist manufacturers. However, in recent years, a number of papers have been published that shed light on this subject. It was generally accepted that the diazoquinone moiety was crucial to the inhibition effect. This belief was supported by the number of chemical reactions reported in the chemical literature between diazoquinone moieties and phenolic derivatives. However, hard evidence had been lacking linking any of these reactions with the inhibition effect. Hard evidence that strong resin–sensitizer interactions are important to the inhibition effect was provided by Beauchemin et al.[20] They showed that the ratio of the infrared peak intensities at 2118 cm^{-1}, P_1; and 2159 cm^{-1}, P_2; can be used to measure the molecular interaction between the resin and the sensitizer. The reasoning is that the stronger the interaction between the resin and the sensitizer, the lower the energy, and the further the shift of the infrared spectra to the red. Furthermore, they



Two-dimensional structure of an *ortho*, *para*'-coupled m-cresol novolac trimer



The three-dimensional representation of an *ortho*, *para*' -coupled m-cresol novolac trimer

FIGURE 19.13 Comparison between the two- and three-dimensional structure of an *ortho*, *para*'-coupled novolac trimer. In this case, the OH groups are pointing outward, which leads to strong intramolecular, instead of intermolecular, hydrogen bonding.

showed that this ratio, P_2/P_1, can be correlated to the dissolution inhibition ability of the sensitizer: the stronger the resin–sensitizer interaction, the stronger the dissolution inhibition. This is in very good agreement with the Domain and the host-guest theories explained in the preceding section.

The hypothesis that the ability of the sensitizer to inhibit the dissolution process is related to the extent to which it interacts with the resin is further explored by Uenishi et al.[21] Like Beauchemin et al. they correlate the shift to the red of infrared spectra of resin–sensitizer mixtures with the dissolution inhibition ability of the sensitizer. Comparing a series of inhibitor structures, they reached the conclusion that the inhibition-ability of the sensitizer is enhanced by keeping the DNQ functional groups in the sensitizer molecule as far apart as possible. The explanation for this observation is that DNQ groups in close proximity compete for the same hydrogen-bonding sites of the resin, weakening the strength of the interaction. They also conclude that the number of DNQ groups should be kept to a minimum, which contradicts early results from other authors.[22]

A key point made by Uenishi et al. is the correlation between the sensitizer hydrophobicity and its dissolution inhibition ability. The parameter they use to measure inhibition is the ratio Rn/Rp, where Rp is the dissolution rate of the resin–sensitizer mixture, and Rn is the dissolution rate of the resin by itself. The retention time in a reverse-phase high-pressure liquid chromatography column was used to estimate the hydrophobicity of the sensitizer. Figure 19.15 and Figure 19.16 show a correlation between the hydrophobicity of dissolution inhibitors and the dissolution inhibition parameter Rn/Rp; this correlation holds true for two different developer concentrations. Furthermore, they show that molecules without the diazoquinone derivative can be effective dissolution inhibitors as long as they are strongly hydrophobic in nature. It is important to clarify that, presently, we are focusing only on the dissolution inhibition properties of the sensitizer. We are neglecting other properties that the sensitizer must have, in order to get incorporated into a photoresist formulation, like adequate sensitivity to the actinic wavelength and chemical compatibility with the resin and solvent.

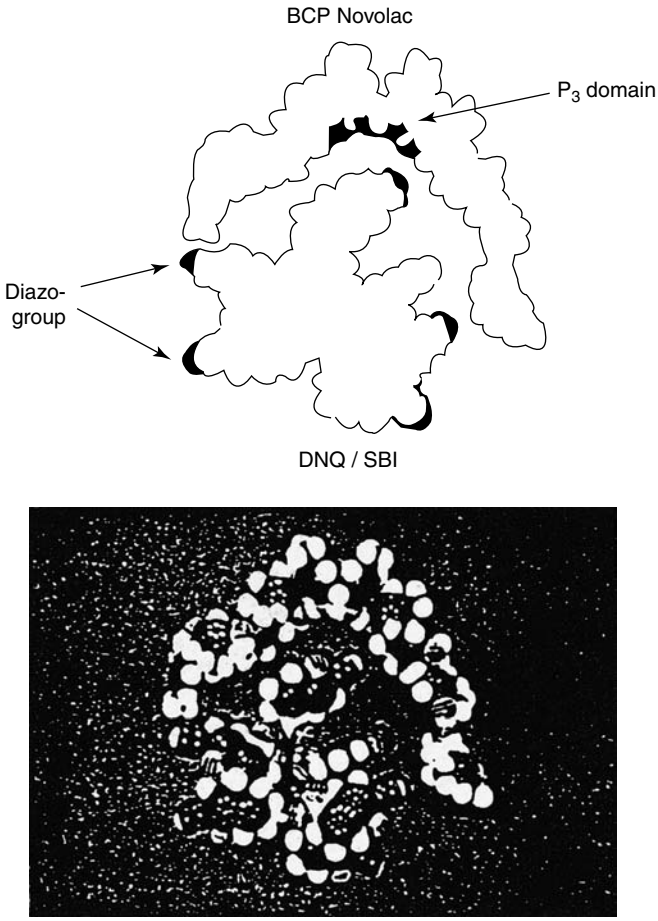


FIGURE 19.14 Molecular models representing the resin–sensitizer interactions according to the domain theory. (From Pawloski, A. R., Torres, J. A., Nealey, P. F., and de Pablo, J. J., *J. Vac. Sci. Technol. B*, 17, no. 6 1999.)

The statement made by Uenishi et al. that inhibition is possible without a diazoquinone moiety is in agreement with an earlier paper from Murata et al.[23]

Notice from these data that there is no difference in the dissolution inhibition produced by structures I and II, despite the fact that structure II does not have a diazoquinone moiety. Furthermore, without the hydrophobic group SO_2Cl, the inhibition effect of structure III has degraded. Finally, structure IV, which contains a diazoquinone group, not only does not inhibit the dissolution process, but also it actually enhances it. This can only be explained in the light of the statement made by Uenishi et al. that within the sensitizer molecule, certain positions of the diazoquinone moiety allow for stronger interaction with the resin than others. Thus, the presence of the diazoquinone moiety is not sufficient for inhibition to take place; there must be an interaction with the resin, and the stronger the better.

In summary, we now know that the dissolution inhibition ability of the sensitizer is affected by:

1. The position and number of DNQ functional groups. The presence of the diazoquinone moiety is not sufficient; its position must be such that strong interactions with the resin take place.
2. Hydrophobic groups must also be present in the structure of the dissolution inhibitor.

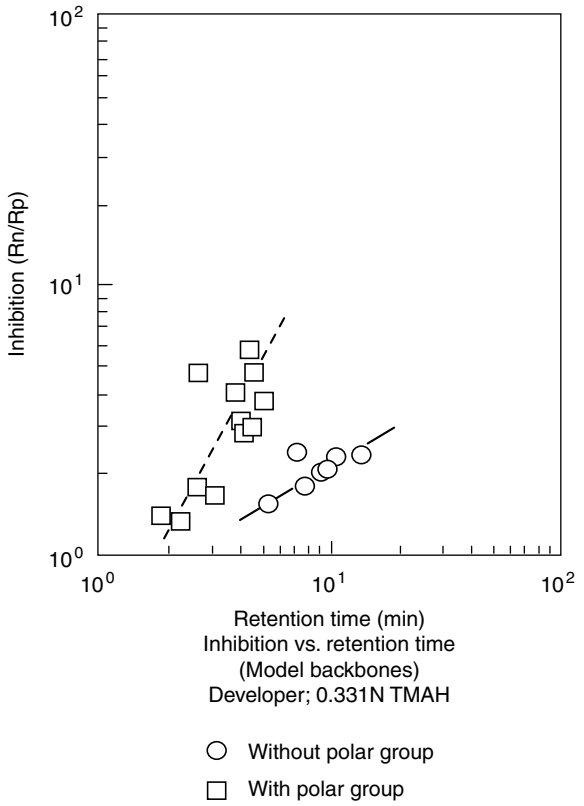


FIGURE 19.15 Dissolution inhibition vs. sensitizer hydrophobicity for novolac resins developed in 0.331 N tetramethylammonium hydroxide (TMAH). The dissolution inhibition effect is measured by the parameter Rn/Rp, where Rp is the dissolution rate of the resin–sensitizer mixture, and Rn is the dissolution rate of the resin by itself. The sensitizer hydrophobicity is measured by the retention time in a reverse-phase high-pressure liquid chromatography column. (From Borzo, M., Rafalko, J. J., Joe, M., Dammel, R. R., Rahman, M. D., and Ziliox, M. A., *Proc. SPIE*, 2438, 294 1995.)

19.2.2.3 Developer

Novolac resins are soluble in strong bases. The first developers were buffered solutions of alkaline bases, typically sodium hydroxide, NaOH. However, when it was discovered that the alkaline metals have an adverse effect on the reliability of semiconductor devices, aqueous solutions of tetramethylammonium hydroxide (TMAH) began being used instead. The chemical formula of TMAH is $N(CH_3)_4OH$, but it is known in the semiconductor industry as TMAH. Today, TMAH-based developers, also called metal-ion-free developers because they do not contain any alkaline cations, are by the only developers used in wafer fabs. Nevertheless, metal-ion developers receive wide coverage in this section. The reason is that they must be included to determine trends important in understanding the mechanism of the dissolution process.

For novolac resins to dissolve at a measurable rate, a minimum base concentration is required, around a pH value of 12.5. This is a clear indication that dissolution of the resin cannot take place without a significant degree of deprotonation. Above that critical concentration, the dissolution rate of exposed and unexposed photoresists increases rapidly as the concentration of the base increases.[24] This holds true for alkaline or TMAH-based developers. Figure 19.17 shows a plot of the dissolution rate of unexposed and exposed photoresist when NaOH is used as developer; and Figure 19.18 plots the dissolution rate for unexposed photoresist using a TMAH-based solution as developer.

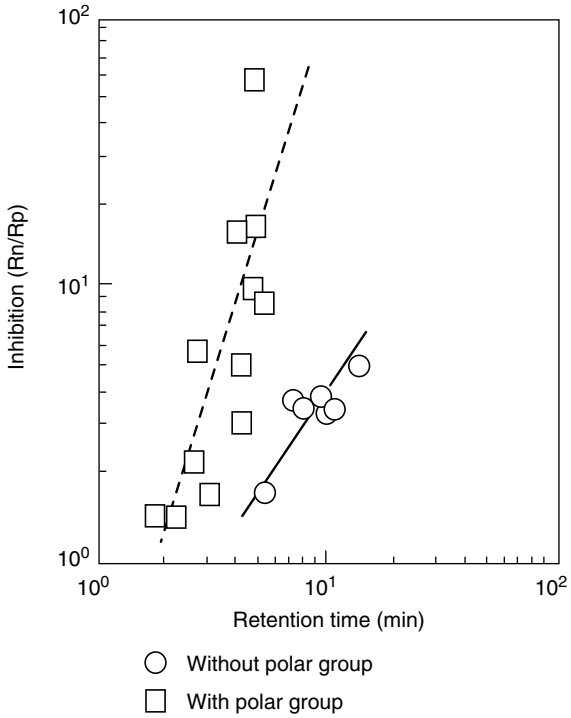


FIGURE 19.16 Dissolution inhibition vs. sensitizer hydrophobicity for novolac resins developed in 0.262 N TMAH. The dissolution inhibition effect is measured by the parameter Rn/Rp, where Rp is the dissolution rate of the resin–sensitizer mixture, and Rn is the dissolution rate of the resin by itself. The sensitizer hydrophobicity is measured by the retention time in a reverse-phase high-pressure liquid chromatography column. (From Borzo, M. J., Rafalko, J., Joe, M., Dammel, R. R., Rahman, M. D., and Ziliox, M. A., *Proc. SPIE*, 2438, 294, 1995.)

Hinsberg et al. found that the dissolution rate of some commercial photoresists developed with NaOH can be described by the empirical equation shown below:

$$\text{Rate} = 1.3 \times 105[\text{Na}^+][\text{OH}^-]^3 \tag{19.2}$$

For exposed photoresist, the dissolution rate's dependence on the concentration of the base is much more complex. Figure 19.18 shows a log–log curve of the dissolution rate of exposed photoresist as a function of $[\text{OH}^-]$. In it we see that, for exposed photoresists, the rate-enhancing effect with increasing [OH] eventually levels off.

As pointed out by Reiser,[25] the rate of dissolution can be expressed in the general form:

$$\text{Rate} = k[\text{cation}^+]^m[\text{OH}^-]^n \tag{19.3}$$

where the exponents m and n are formal reaction orders. This implies that the OH anion and the corresponding cation are involved in the rate-determining step of the dissolution process.

Hinsberg et al. and Huang et al.[26] have studied the effect of the nature and size of the base cation on the dissolution rate. Figure 19.19 shows the dissolution rate of *para*-nitrosubstituted novolac resins developed with 0.08 N solutions of bases of different monovalent cations. Table 19.2 lists the radii of hydrated and unhydrated alkali ions. As it can clearly be seen in Figure 19.20, the dissolution rate correlates with the radii of the unhydrated, not the hydrated, cation.

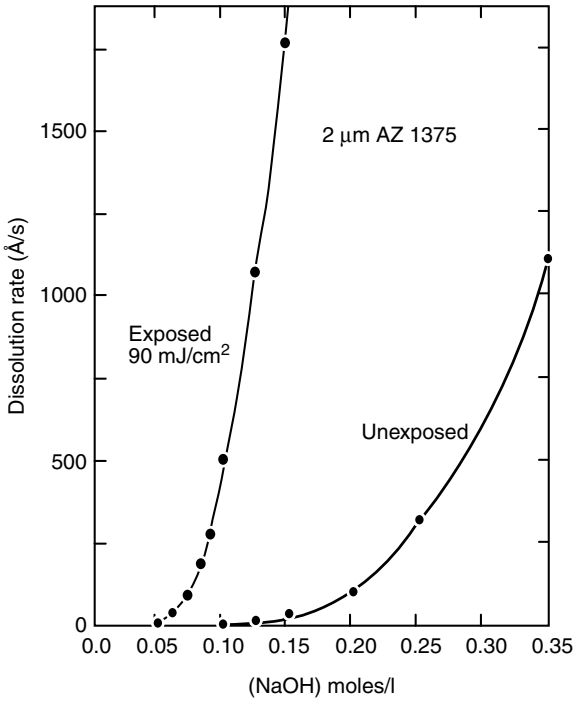


FIGURE 19.17 Dissolution rates of exposed and unexposed AZ1375 photoresist as a function of developer sodium hydroxide concentration. (From Trefonas, P. III, and Daniels, B. K., *Proc. SPIE*, 771, 194,1987.)

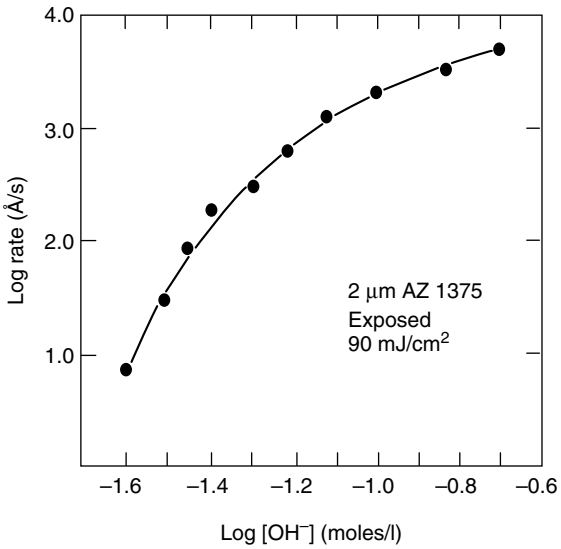


FIGURE 19.18 Dissolution rate of exposed AZ1375 photoresist as a function of developer hydroxide ion concentration. (From Trefonas, P. III and Daniels, B. K., *Proc. SPIE*, 771, 194, 1987.)

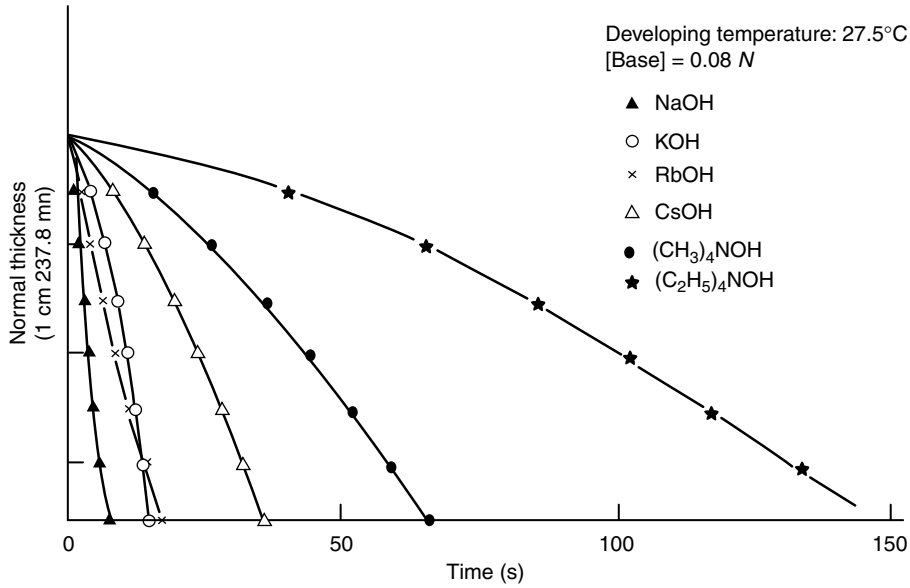


FIGURE 19.19 Effect of the cation of the developer base. Dissolution curves (resist thickness vs. time) of an experimental phenol-novolac resin in 0.08 M solution of the hydroxides indicated in the figure. (With permission from W.D. Hinsberg, et al., personal communication.)

The trends shown thus far have applied to metal-based or metal-ion-free developers equally. One important difference between the two types of developers, however, is the change in the dissolution rate as a function of temperature. Metal-ion-based developers show an increase in the dissolution rates as a function of temperature, as we can see in Figure 19.21. This is the expected trend from the Arrhenius Law:

$$k = Ae^{-(E/RT)} \tag{19.4}$$

where k is the chemical reaction rate constant, A is a proportionality constant related to the activation entropy, E is the activation energy, R is the universal gas constant, and T is the absolute temperature. Metal-ion-free developers, however, seem to follow the opposite trend, as shown in Figure 19.22.

In summary, the concentration, chemical make-up, and temperature of the developer have a great impact on the dissolution rate:

1. The dissolution rates for exposed and unexposed photoresist increase with the concentration of the base and pH.
2. The dissolution rates decrease with the radii of the unsolvated base cation.
3. The dissolution rate increases with temperature for alkaline-base developers but decreases (within a certain range) for metal-ion-free developers.

TABLE 19.2 Radii of Hydrated and Unhydrated Alkali Ions (Å)

	Li^+	Na^+	K^+	Rb^+	Cs^+
Crystalline Radii	0.68	0.98	1.33	1.48	1.67
Hydrated Ions	3.40	2.76	2.32	2.28	2.28

Source: From Huang, J. P., Kwei, T. K., and Reiser, A. *Proc. SPIE* 1086, 74, 1989.

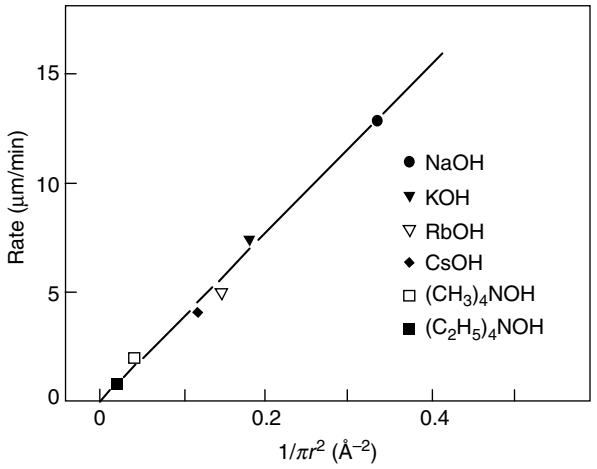


FIGURE 19.20 Dissolution rates from blends of *para*-nitrosubstituted novolac with 10% by weight of poly (2-methylpentene-1-sulfone) (PMPS) in hydroxide solutions of different cations (0.08 N, 27.5°C). The rate is plotted as a function of the reciprocal of the cation cross-section, calculated from crystallographic data. (With permission from W. D. Hinsberg, et al., personal communication.)

Any mechanism proposed to explain the dissolution of novolac resins and formation of relief images must be consistent with these observations.

19.2.3 Development Mechanisms in Novolac-Based Photoresists

Ueberreiter and Asmussen[27,28] studied the dissolution of high-molecular-weight polymers in solvents. They found that the dissolution process takes place in two stages:

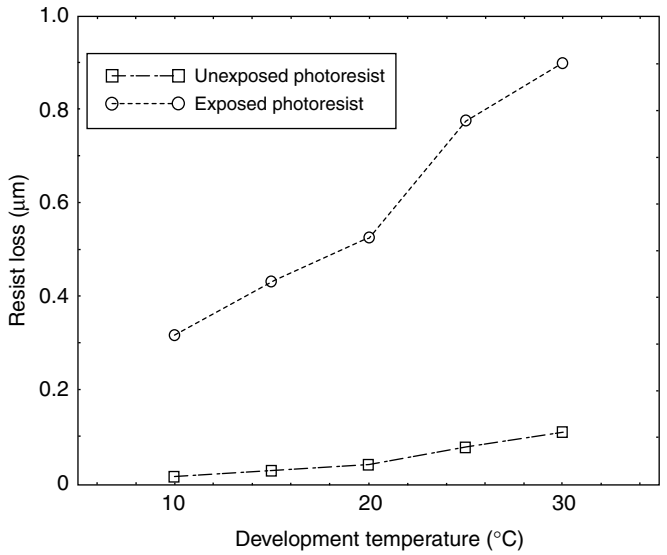


FIGURE 19.21 Relative rate of exposed and unexposed photoresist as a function of development temperature for a metal-ion base developer.

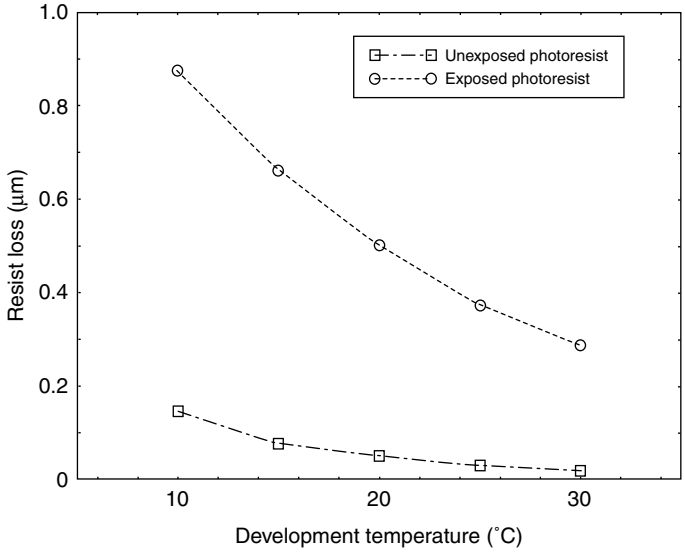


FIGURE 19.22 Relative rate of exposed and unexposed photoresist as a function of development temperature for a metal-ion-free base developer. (From Reiser, A., *Photoreactive Polymers*, WileyInterscience: New York, 1989, 179.)

1. In the first stage, the solvent penetrates the glassy structure of the polymer. As a result of this penetration, a gel layer that separates the polymer's solid phase from the solvent's liquid phase is formed.
2. In the second stage, the polymer coils disentangle and diffuse into the liquid phase of the solvent. Thus, in the dissolution steady state, three phases are present: the glassy polymer, the swollen gel, and the polymer in solution.

The dissolution of novolac resins is not the result of mixing a polymer with a compatible solvent, but rather the result of a chemical reaction that forms a product with a large affinity for the developer. Nevertheless, Ueberreiter and Asmussen's analysis can be very useful to U.S. in determining where the rate-determining step is taking place.

In the development of high-molecular-weight novolac resins, a swollen gel has also been observed. If the rate-determining step is the diffusion of the developer across the gel layer, the dissolution process will follow Fick's Law, and the rate will depend on the square root of time. If, on the other hand, one of the events taking place at the polymer–gel or gel–developer interface is the rate-determining step, the development process will be a linear function of time. This is called Case II dissolution, or polymer-relaxation-controlled mass transfer.[29]

The experimental data clearly show that the dissolution rate of novolac resins does not follow the square root of time. Rather, the dissolution rate is a linear function of time. The only exceptions are changes in the dissolution rate caused by standing waves formed during the exposure process, or inhibition's effects at the resist–surface or resist–substrate interface. This clearly points out to one or several of the events at the resist-gel or gel-developer interface as being the rate-determining step.

The steps involved in the dissolution process are:

1. Diffusion of OH^- ions and water into the polymer matrix.
2. Deprotonation of phenol groups of the novolac resin to polymer-bound phenolate ions.
3. Solvation of the phenolate ions.

4. Compensation of the negative charge of the phenolate ions by the positive charge of the base cation and formation of ion pairs.
5. Rearrangement of the ionized polymer chains, detachment from the polymer matrix, and transfer into solution.

We just determined that the diffusion of the OH^- ions is not the rate-determining step because the development rate does not follow Fick's Law. This leaves steps 25 as candidates for the rate-determining step.

The ionized polymer chains clearly will be more stable in solution than in the organic polymer matrix. Since, there are no steric effects constraining this process, it is likely to take place very rapidly. This excludes step 5 and leaves steps 24 as the candidates for the rate-determining step.

The four major models proposed in the literature to explain the dissolution process of novolac resins in basic aqueous solutions are the Membrane Model, the Secondary Structure Model,[30] the Percolation Model, and the Critical-Ionization Model.[31] All four of them focus on one or more of steps 24 as the rate-determining step or steps.

19.2.3.1 The Membrane Model

In the Membrane Model, Arcus views the interface between the novolac-resin matrix and the developer as a membrane that "… can differentiate between the ions of aqueous basic developers due to variations in size, composition, and charge." This membrane postulated by Arcus would let the OH ions pass but slow down the larger cations. Thus, the transport rate of the cation, required to stabilize the phenolate anion by the formation of ion pairs, will be the rate-determining step.

This model can explain some experimental observations, like the fact that adding neutral salts of the cation to the developer solution enhances the dissolution rate. However, it is in disagreement with some others. For instance, this model predicts a decrease of the dissolution rate with the hydrated size of the cation, whereas the correlation is with the size of the unhydrated cations. More important yet, it fails to explain the very large observed differences in the dissolution rate of isomeric resins.

19.2.3.2 The Secondary Structure Model

A single person did not propose the secondary structure model, but rather it evolved over time. The principal players are, Templeton, Szmanda, Trefonas, Daniels, Garza. This model proposes that the rate-determining step is the deprotonation of the resin, step 2 above. Furthermore, it stresses the role of the three-dimensional structure of the resin during the deprotonation reaction. Certain resin configurations, like those with predominant *para–para* bonding, have –OH groups more easily accessible to deprotonation and will dissolve at a higher rate. Also proposed is the possibility that, in the case of resins with prevalent *ortho–ortho* bonding, the close spatial proximity of the OH resin groups can stabilize a phenolate ion by distributing the negative charge over a larger area, as shown in Figure 19.23.

This stabilization will reduce the likelihood of further deprotonation, keeping the phenolate ion in the polymer matrix, and reducing the dissolution rate. This model, however, is incomplete because it fails to explain many experimental observations, in particular, the role of the sensitizer, such as the large dependency of the photoresist dissolution rate on the hydrophobicity of the sensitizer.

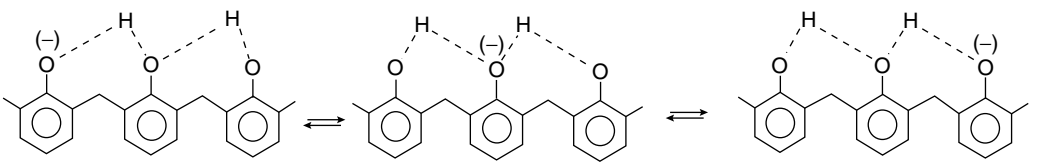


FIGURE 19.23 Stabilization of the phenolate ion by distributing the negative charge over a number of OH groups through hydrogen bonding. (From Yeh, T.-F., Shih, H.-Y., and Reiser, A. A., *Macromolecules*, 25, 5345, 1992.)

19.2.3.3 The Percolation Model

The Percolation Model by Reiser and co-workers[31,32] is by far the most developed theory describing novolac dissolution. It has been used to describe many of the factors affecting the dissolution rate of novolac: dissolution inhibitors[33] exposed resist films,[34] dissolution promoters,[35] added salts,[36] the base cation,[37] isotopic substitution,[38] and resin molecular weight.[39] The Percolation Model borrows the concept of the transition zone (or gel layer) from the Membrane Model of Arcus and adapts percolation theory[40] to describe the diffusive properties of the gel layer. In the Percolation Model, the gel layer, also referred to as the penetration zone, forms when the concentration of phenolate groups in equilibrium with the developer at the interface between the developer and the resin reaches a limit of solubility. The penetration zone, therefore, is a distinct polymer phase, having its own T_g, and the developer can diffuse much faster in the penetration zone than in the bulk polymer matrix. If the developer concentration does not exceed a critical minimum value, the penetration zone does not form, and dissolution occurs at a rate that is several orders of magnitude slower than regular development.

This theory describes a penetration zone that grows thicker as developer continues to transfer into the zone from the bulk solution. The diffusion of base within the zone and the reaction of base with the matrix resin lead to a gradient in the phenolate concentration. When the degree of conversion of phenol to phenolate at the back end of the zone reaches a critical value, the novolac chains dissolve. At this point, the penetration zone reaches a constant thickness, and dissolution continues in a steady state.

The rate at, which the penetration zone travels is identically the rate at, which the novolac film dissolves. The Percolation Model suggests that the diffusive flux of base through the penetration zone is the product of the velocity of the boundary and the mean base concentration in the zone. Huang et al. showed that the diffusion coefficient drops precipitously at the interface of the zone and the solid matrix, indicating that the rate-limiting step occurs at this interface. The strong dependence of the dissolution rate on cation size led these authors to conclude that the rate-limiting step in the dissolution process is the disassociation of the developer cations from their hydration shells. This statement was the first definitive pronouncement concerning the rate-determining step according to the Percolation Model. The Percolation Model currently emphasizes the deprotonation reaction and transfer of cations at the front of the penetration zone and attaches less significance to the diffusion of base through or across the penetration zone.

In the Percolation Model, the propagation of base proceeds through channels made of hydrophilic sites, identified as the phenol groups or, alternatively, as the phenolate ion pairs, dispersed in a matrix of hydrophobic material. Dissolution occurs only when the hydrophilic sites are sufficiently concentrated to form a continuous network. The fraction of neighboring sites open to the propagation of base is called the percolation parameter, p, and is linearly related to the fraction of "occupied cells," e.g., the fraction of base-accessible phenolic repeating units in amphiphilic copolymers. The percolation threshold, pc, is the value of the percolation parameter below, which dissolution does not occur. The dissolution rates of many copolymers have been shown to be proportional to $(p - \mathrm{pc})^2$ where, $\mathrm{pc} = 0.2$.

The Percolation Model explains most effects on the dissolution rate by corresponding changes in the value of the percolation parameter, which reflects the relative difficulty in attaining continuous channels for percolation. Polyhydroxystyrene, which has a continuous spiral of hydroxyl groups along its backbone, has an unbroken succession of these hydrophilic sites, whereas the hydroxyl groups in *ortho–ortho* linked *p*-cresol novolac aggregate in clusters, thereby breaking the hydrophilic pathways. Hydrophobic additives will affect the dissolution rate only if they interact with the hydrophilic sites, while hydrophilic additives normally increase the dissolution rate because they usually add hydrophilic sites that facilitate percolation.

According to this theory, dissolution inhibitors operate by effectively disabling the hydrophilic sites and obstructing pathways for percolation. The hydrophobic displacement volume of the inhibitor depends not only on the structure of the inhibitor, but also on the mobility of the resin. The ability of a single inhibitor molecule to disable as many as 16 hydroxyl groups at a time led Shih and Reiser to propose that multiple sites are removed from the percolation pathway through the inhibitor's inductive

polarization of hydrogen-bonded clusters of hydroxyl groups. The inhibitor, therefore, disrupts the otherwise random distribution of percolation sites by lowering the site connectivity and increasing the T_g of the penetration zone. If dissolution promoters (accelerators) are present, the dissolution rate depends on whether or not the accelerators are included in the phenolic clusters, which are induced by the inhibitor. The acceptance of the accelerator into a cluster is dependent on its acidity relative to that of novolac. At low inhibitor concentrations, the accelerators, which do not compete well with the more acidic novolac remain outside the clusters and increase the concentration of percolation sites. At higher inhibitor concentrations, the accelerators are accepted into the phenolic clusters, and their effect on the dissolution rate is diminished.

Shih and Reiser suggest that the dissolution promotion that occurs upon exposure is concomitant with an elimination of phenolic clusters. The heat released during the Wolff rearrangement of the DNQ photolysis product releases the polymer chains in the vicinity of the reacting DNQ, thereby dispersing the cluster. The carboxylate ions that appear in the penetration zone add hydrophilic sites to the percolation field and lower the T_g of the penetration zone.

19.2.3.4 The Critical-Ionization Model

The Critical-Ionization Model[41] provides an understanding at the molecular level of the important factors in the aqueous dissolution of phenolic polymers below the entanglement molecular weight. The model postulates that a critical fraction of the acidic sites on a phenolic polymer must ionize for the polymer to dissolve in aqueous base. A functional relationship between the dissolution rate and the degree of ionization was developed based on this hypothesis. The model provides an explanation for the critical-base-concentration phenomenon and for the dependence of rate on molecular weight, phenomena that are not readily explained by other models. Quantitative predictions for the effects of polymer structure on the dissolution rate follow from Equation relating the degree of ionization to the degree of polymerization, the polymer p*K*a, and the developer concentration.[42] Experimental verification has been provided through tests of model predictions for the minimum base concentration required for development and the effects of polymer structure on the dissolution rate.

Molecular simulations of resist dissolution based on the Critical-Ionization Model[43,44] were used to probe the mechanism of surface inhibition and the evolution of edge roughness and surface roughness in photoresist profiles. These simulations demonstrate the dependence of the dissolution rate and surface roughness on the molecular-weight distribution of the polymer, degree of deprotection, void fraction, and developer concentration. Model parameters were evaluated using experimental data from turbidimetry, potentiometry, and copolymer studies.

19.3 Formation of the Relief Image in Chemically Amplified Resists

19.3.1 Overview

As explained in Section 1.1, the path the industry has followed to improve resolution has been to reduce the actinic wavelength once the practical limit for increasing the NA has been reached. By the early 1980s, there was a major industrial effort underway to develop photoresist systems at shorter wavelengths[45] than 365 nm. The next readily available wavelength is in the deep UV (DUV) region near 250 nm. Two sources exist in this region: the mercury discharge lamp has a small emission peak centered at 254 nm, and a stronger source from a krypton-fluoride, KrF, excimer laser is available at 248 nm. Other sources at shorter wavelengths were proposed, and resist systems were developed to work at these wavelengths as well.[46–48]

Conversion of the DNQ/Novolac platform to DUV exposures faced several challenges. The materials used in the DNQ/Novolac platform are too highly absorbant at 250 nm to obtain vertical profiles. Furthermore, the intensity from the available light sources and increasingly complex-imaging systems is two orders of magnitude smaller than the 365-nm systems. The Novolac resin itself is very opaque at

250 nm (ca 0.5 μm^{-1}). Many replacement resins with better optical properties were developed and tried over a 10-year period, Figure 19.22. These ranged from isomerically pure novolac,[49] pHOST,[50] pHOST copolymers,[51] and acrylate polymers.[52] Also the DNQ dissolution inhibitor compounds have too high of an unbleachable absorbance at 250 nm. This limits the ultimate loading of the photoactive compound's (PAC) and the performance of DNQ resists in the DUV region. To overcome this limitation, other PAC inhibitors were developed, Figure 19.23. These included Meldrum's diazo and related compounds from IBM[13] and BASF,[53] *meta* nitrobenzene compounds from AT&T Bell Labs,[54] and bis(arylazides) from Hitachi.[55] However, none of these were proven to be commercially successful.

The initial attempts at DUV wavelength photoresists focused on either chain scissioning of poly methyl methacrylate (PMMA) polymers[56,57] or extension of the DNQ/Novolac platform to work with 248-nm exposure. Chain scissioning resists of PMMA polymers are capable of extremely small resolution. Unfortunately, they suffer from several drawbacks. First, the photospeed of such materials is relatively large. Also, the same properties that allow the polymer chain to scission decrease its etch resistance. Finally, an organic developer is required to image these resists, which is undesirable from an environmental/regulatory standpoint.

The most difficult challenge in developing DUV photoresists is the extremely low output of the mercury discharge lamp at 254 nm. Because the number of photons available in the DUV region from this source is approximately 100 times fewer than the 365-nm i-line source, a comparable increase in photospeed is required to maintain acceptable wafer throughput levels. Though significant advances were made in the conventional dissolution inhibitor system photospeed, nothing approaching two orders of magnitude improvement was found. A new photoresist design was clearly needed. Researchers at IBM pioneered the concept of chemical amplification to achieve the desired photospeed for DUV resist systems.[58] For chemically amplified systems, the exposure step generates a catalyst in the film. During a subsequent thermal bake step, this photogenerated compound catalyzes chemical change to the resin or additives to influence the solubility of the resist matrix. Because the photogenerated catalyst is not lost in the basic reaction it can continue reacting with the resin many times. A single photocatalyst can affect several hundred reactions within the film permitting the needed photospeed enhancement.

IBM simultaneously pursued three classes of chemically amplified resists: crosslinking,[59] depolymerization,[60] and site deprotection. The crosslinking systems used a photogenerated Lewis acid to initiate cationic polymerization of epoxide side chains. The crosslinked areas become insoluble and the resist works effectively as a negative imaging system. Several other groups developed crosslinking type chemically amplified resists.[61,62] Some have been commercialized and are in use today.

The depolymerization system makes use of the low ceiling temperature of certain polymers. Any polymer heated above its ceiling temperature will depolymerize into monomeric components. This depolymerization starts from the ends of the polymer and the long chain essentially unzips one monomer at a time. IBM made use of this phenomenon by capping low ceiling temperature polyphthaldehyde polymers with thermally stable end groups. Photogenerated acids could, however, break the polymer at one of the repeating units. The polymer then unzips to highly volatile units and essentially self develops. These systems exhibit very poor etch resistance because the polymer unzips under plasma etch conditions, and they have been essentially abandoned.

The third chemically amplified system pursued has ultimately proven to be the most useful. Base soluble sites (phenols, carboxylic acids, etc.) can be capped with protecting groups making them insoluble. Under certain conditions, these protecting groups can be catalytically removed to return the solubility in aqueous base. Several protection schemes have shown great utility in chemical synthesis. IBM developed a photoresist in which phenolic sites on (pHOST) have been protected with *t*-butoxycarbonate groups.[63] At extreme temperatures ($>$180C), the protected sites thermolyze and yield gaseous CO_2, iso-butylene, and deprotected pHOST. In the presence of a strong acid, this thermolysis happens at much lower temperatures ($<$90C). A post-exposure bake (PEB) at 90°C selectively cleaves protecting groups only in areas of photogenerated acid. The IBM tert-butyloxy carbonates (*t*-BOC) photoresist works in positive tone using aqueous developer and alternately in negative tone using an organic developer.

The process flow for chemically amplified resists is the same as for conventional resists (see Figure 19.3). The underlying chemistry, however, is different for several of the steps in the flow. These differences have dramatic implications for processing and the level of control needed when manufacturing semiconductor devices with chemically amplified resists. For instance, the PEB, for conventional resists is designed to remove standing wave patterns from the exposure step. The PEB step can be considered a secondary control on i-line resist performance. For chemically amplified resists the PEB step is necessary to drive the catalytic reaction to completion. The PEB step, therefore, has primary influence on DUV resist performance. The temperature stability and uniformity requirements of the wafer bake plates are much more demanding for chemically amplified resists than for conventional resists. For instance, typical 193-nm resist systems show 0.5–3 nm critical dimension (CD) change for a 1°C temperature variation.

In the next few sections the individual processing steps will be covered for chemically amplified resists. The detailed theory and mechanisms will be given for each step. We begin with the exposure step and generation of the catalytic photoacid.

19.3.2 Exposure Step

As with conventional resists the exposure step converts a neutral molecule into an acid product. For conventional resists the final photoproduct is a carboxylic acid. And although the resist companies would incorporate the DNQ group in many different forms, the basic mechanism for this photoreaction is the same for the vast majority of DNQ/Novolac systems. Typically with chemically amplified systems, the acid strength of the photoacid (Photo Acid Generator PAG) is significantly higher than a carboxylic acid and many new PAGs have been developed.

All PAG molecules can be described as the acid counter ion integrated with a photoactive ballast group. The ballast decomposes upon exposure and releases the counterion. The acid proton can either come from the decomposed ballast group or be extracted from the host matrix.

(19.5)

Each PAG can be classified by the nature of the counter ion (or liberated acid) and the chemical structure of the ballast group. The original PAGs used by Ito et al.[64] were based upon very strong or superacids, $HSbF_6$ and $HAsF_6$. These superacids worked well for the chemical amplification and were readily synthesized, but posed a device contamination risk due to the inclusion of the heavy elements Arsenic and Antimony. Subsequent PAGs have generally been based upon sulfonic acids such as tosylates. The acid strength and size of photoacids have varied greatly from system to system. Table 19.3 gives a

TABLE 19.3 Corrected C Parameters, Quantum Yields of Acid Generation, Normalized

PAG	Matrix	PAG Loading	Exposure 1	Films Abs mm^{-1}	Corrected C $cm^2\ mJ^{-1}$	Facid Total	Eo mJ cm^{-2}
TPS	Phenolic	Low	248	0.241	0.055	0.27	8.7
DTBPI	Phenolic	Low	248	0.25	0.057	0.28	8.1
TPS	Acrylate	Low	248	0.1	0.047	0.63	4.8
DTBPI	Acrylate	Low	248	0.111	0.018	0.22	12.2
TPS	Acrylate	Low	193	0.538	0.029	0.11	7.2
DTBPI	Acrylate	O	193	0.496	0.012	0.05	20.1
TPS	Acrylate	High	193	0.885	0.026	0.13	7.3
DTBPI	Acrylate	High	193	0.829	0.008	0.04	23.2
TPS	Phenolic	High	248	0.299	0.042	0.33	5.2
DTBPI	Phenolic	High	248	0.32	0.051	0.39	4.8

Film absorbances and dose to clear for both PAG's under different conditions.
Source: From Courtesy of Jim Cameron-Rohm and Haas Electronic Materials.

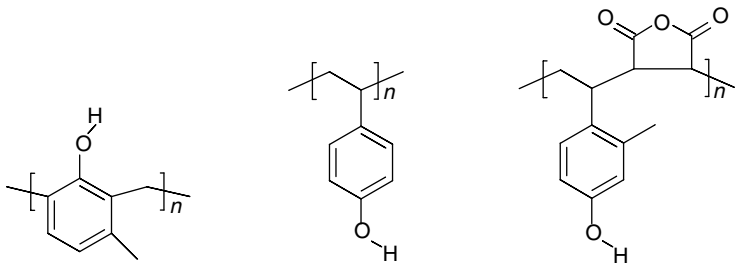


FIGURE 19.24 Potential deep UV (DUV) resist resins.

representative list of photogenerated acids used in resist systems along with some relative physical data. The choice of the counter ion used in the PAG greatly influences the amplification step, boiling point; molecular weight determine volatility and *t*-topping, Van der Waals volume determines diffusion and iso-nested bias (define), p*K*a determines deprotection rate and susceptibility to environmental contamination, etc.

Many ballast groups have been designed into PAGs. However, four designs have been most prominent in DUV resists. These are based on onium salts, diazosulfone compounds, nitrobenzyl esters, and sulfonyloxy imides. Representative structures for each class of PAG are shown in Figure 19.24. Because the mechanism for each PAG class is different, we will cover each of these classes separately.

19.3.2.1 Exposure Step Chemistry

19.3.2.1.1 Onium Salts

The most common PAGs in commercial use today are alkyl-onium salts. The onium salt PAGs were first developed by Crivello for use in photocurable epoxy resins.[65,66] The mechanism for acid production has been extensively studied by researchers at IBM[67–70] and is shown in Figure 19.25. Two competing pathways exist with both leading to acid production. Exposure of the PAG heterolytically cleaves the phenyl–sulfur bond generating a phenyl radical and a radical cation centered on the sulfur atom. In-cage recombination liberates the acidic proton from the parent compound through the first pathway. If the radical pair escapes from each other, the radical cation must abstract a proton from the surrounding matrix before dissociating into the acid product. The data is unclear about the source of the proton (relative humidity) in this cage escape pathway, but the polymer matrix certainly plays a role in determining the relative yield of photogenerated acid. A similar dual pathway was observed for iodonium PAGs.[71]

19.3.2.1.2 Sulfonyloxy Imides

The use of sulfonyloxy imide PAGs as photoacid generators was first patented in 1983.[72] It was not until 1990, however, that information about the photomechanism for these PAGs began appearing in the open literature. Direct excitation leads to homolytic cleavage of the N–O bond giving a radical pair

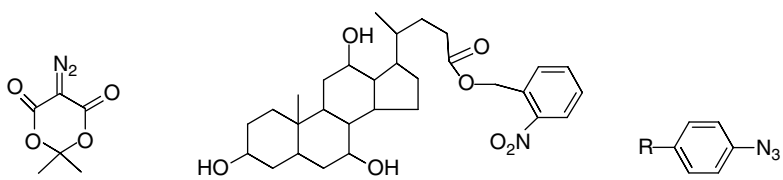


FIGURE 19.25 DUV compatible dissolution inhibitors.

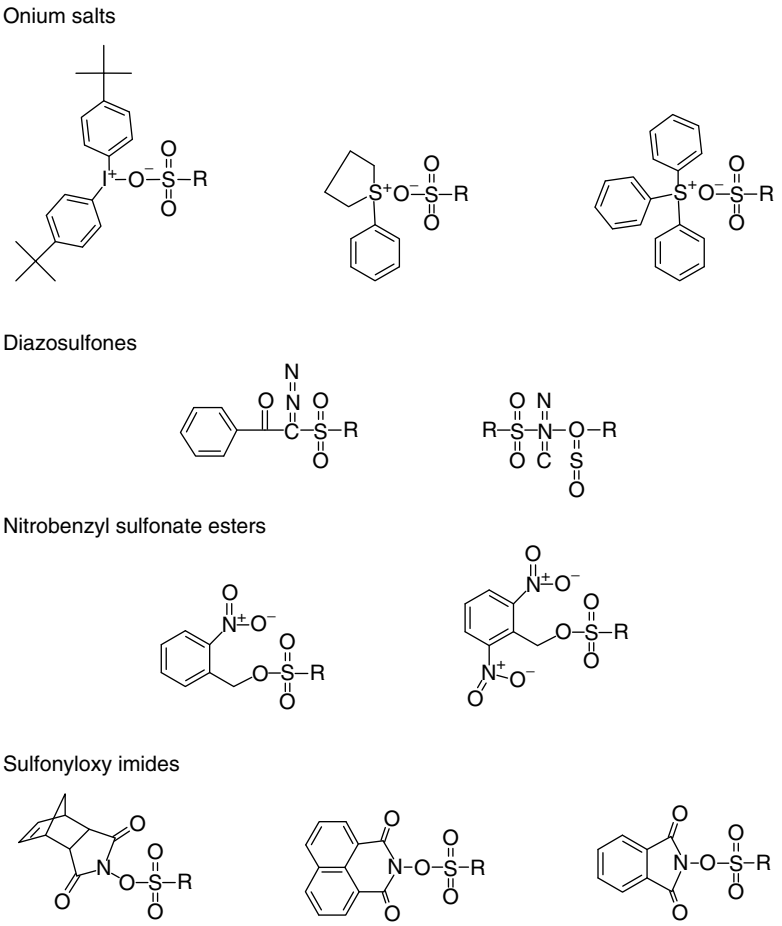


FIGURE 19.26 Representative photoacid generation (PAG) structures.

(Figure 19.26). If cage escape occurs rapidly the sulfonyl radical can abstract a hydrogen from the matrix to produce the desired photoacid. In-cage recombination returns the original PAG structure limiting the quantum yield. Alternately, the π imidyl radical can undergo ring opening.[73] Recombination with the sulfonyl radical, at this point, will yield a non-acidic photoproduct, although this recombination is reportedly unfavorable.[74] Kasai studied the stability of π imidyl radicals and observed that the ring's ability to open is controlled by the contribution of the π electrons to the ground state. Five-member rings with extended conjugation open more readily, and six-membered rings do not open. The relative quantum yield for acid production measured by Szmanda et al. for various sulfonyloxy imides, PAGs partially confirms this observation.

An alternate photochemical pathway involving photoinduced electron transfer has been proposed by Brunsvold and is shown in Figure 19.27. Here, a donor molecule is excited upon absorption. If the oxidation potential of the excited state donor is lower than the reduction potential of the PAG, then electron transfer is thermodynamically favorable. With the addition of an electron the PAG becomes unstable. Protonation induces decomposition into the sulfonyl radical, which is free to abstract another proton from the matrix to yield the desired photoacid. Photospeed enhancements of 5× were obtained by the addition of sensitizers that have significant absorbance at the exposure wavelength and a low enough excited state oxidation potential.

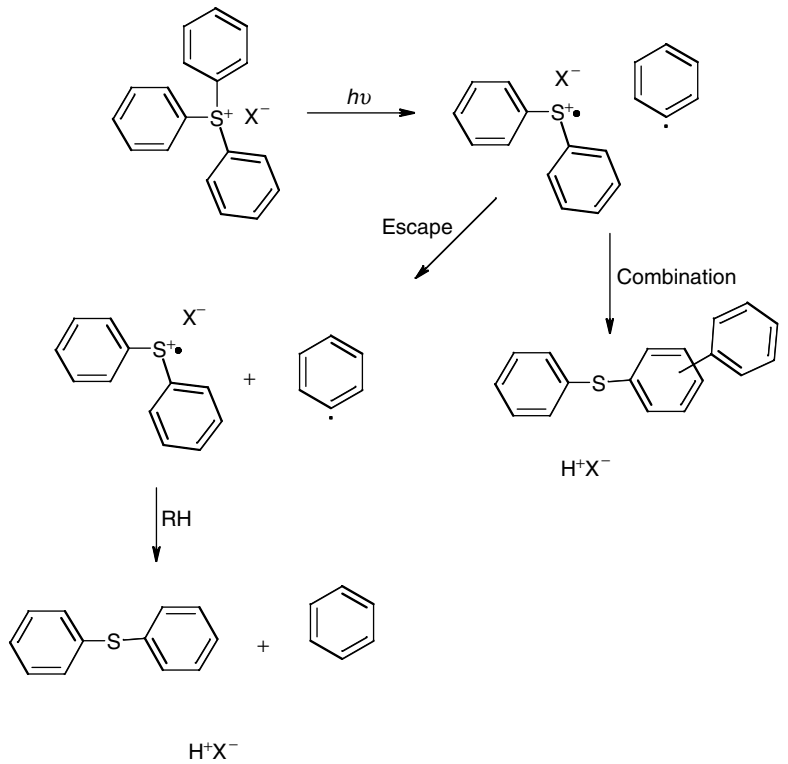


FIGURE 19.27 Reaction pathway for sulfonium PAG decomposition.

19.3.2.1.3 Diazosulfonyl Compounds

The diazosulfonyl PAGs were originally developed as dissolution inhibitors of novolac resins for which they functioned poorly.[75] Researchers at Hoechst saw their potential as PAGs of aromatic sulfonic acids.[76] The photochemical pathway is similar to that of the Meldrum's diazo compound described by Willson et al. Photolysis proceeds by the loss of N_2 upon exposure followed by Wolf rearrangement to yield a reactive sulfene. Subsequent reaction with residual water yields the catalytic acid. Several non-acid photoproducts have been detected, which result from alternate reactive pathways as depicted in Figure 19.28. Through the loss of the diazo chromophore, significant bleaching of absorbance at 248 nm occurs with photolysis of these PAGs. However, because of the high polymer absorbance, low PAG loading, and relatively low conversion at lithographic doses (20% @4.5 mJ), the bleaching contributes little to the lithographic performance of DUV resists. The thermal stabillity of the diazosulfonyl PAGs is lower (Td ~ 160°C) than other PAGs described here. For the acetal based photoresists, for which these PAGs where intended, this thermal stability is sufficient.

19.3.2.1.4 Nitrobenzyl Sulfonate Esters

Reichmanis et al. developed a novel dissolution inhibitor type resist for 248 nm using *o*-nitrobenzyl carboxylate inhibitors. This photodecomposition reaction was, later, used to develop PAGs for chemically amplified resists[77](Figure 19.29). The carboxylic ester functionality was replaced by a sulfonate ester leading to a sulfonic acid photoproduct that is strong enough to act as a deprotection catalyst. The original nitrobenzyl ester PAGs (Ra = H, Rα = H) had poor thermal stability (Td ~ 100°C). The thermal stability could be enhanced by incorporating electron withdrawing groups (Ra = CF_3, NO_2, Cl, etc.) in the *ortho* position on the benzene ring or bulky substituents (Rα = $COCH_3$, $COCH_2CH_3$, etc.) at the α

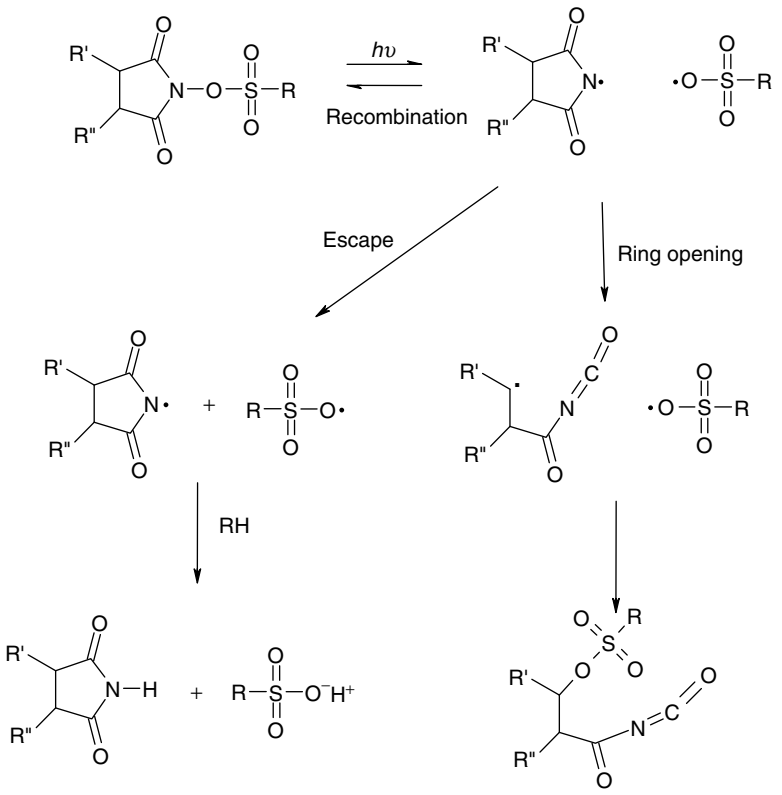


FIGURE 19.28 Direct excitation pathway for sulfonyloxy-imide PAG decomposition.

position.[78] Maximum thermal stability when both tactics are used is well over 250°C. A reduction in quantum yield is observed with incorporation of the bulky substituents. However, the ability to PEB the resist at a higher temperature compensates for the lower quantum yield.

19.3.2.2 Exposure Step Kinetics

The kinetics for acid generation regardless of PAG type has often been described using a first order kinetic model similar to Dill's model of DNQ photodecomposition.[79]

$$\text{PAG} + h\nu \rightarrow \text{Acid} + \text{photoproducts}$$

$$\frac{\text{d[PAG]}}{\text{d}t} = -\,\text{C[PAG]}I_{h\nu} \tag{19.6}$$

$$[\text{Acid}]_{\text{dose}} = [\text{PAG}]_{\text{dose}=0} - [\text{PAG}]_{\text{dose}} = [\text{PAG}]_{\text{dose}=0}(1 - \text{e}^{-\text{C Dose}})$$

The ability of Equation 19.6 to model acid yield vs. dose has been confirmed by many researchers.[80,81] Except for the diazosulfone PAGs, very little photobleaching has been observed since the decomposition products are just as strongly absorbing at 248 nm as the parent PAG. Even with the diazosulfonate PAGs, the loading of PAG and low conversion levels at lithographics doses minimizes the effect of bleaching on DUV resist performance.

Alternate kinetic models for PAG decomposition have been reported.[82] The most prominent involves photosensitization of the PAG molecule by a donor compound within the resist. This is shown in

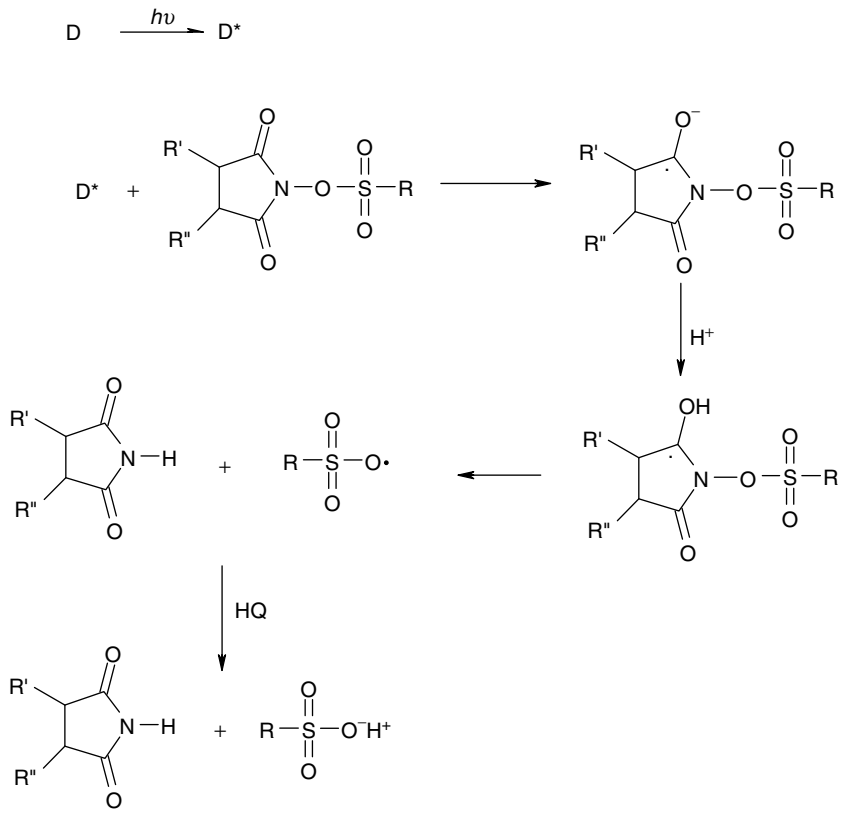


FIGURE 19.29 Photoinduced electron transfer pathway for sulfonyloxy-imide PAG decomposition.

Figure 19.30. The donor molecule is excited by absorption of a photon in the first step. In step 2, the PAG is promoted to an excited state as the donor molecule relaxes back to a ground state. The excited molecule then decomposes to yield an acid product. The donor, D, can be either a chromophore on the polymer or another molecule added to the resist.

The photosensitization depicted in Figure 19.30 can occur by either electronic energy transfer or excited state electron transfer. Electronic energy transfer requires overlap between the emission spectrum of the donor molecule and the excitation spectrum of the acceptor molecule (PAG). This clearly is not the case for many DUV resists in which the excitation spectrum of the PAG is much higher in energy than the polymer emission spectrum. For excited state electron transfer to be allowed, the oxidation potential for the excited state donor molecule must be lower than the reduction potential of the PAG molecule.

$$\Delta G_{\text{transfer}} = E_{\text{ox(D}^*\text{)}} - E_{\text{red(PAG)}} \sim E_{\text{ox(D)}} - E_{\text{(D}^*\text{)}} - E_{\text{red(PAG)}} \leq 0 \tag{19.7}$$

For the case of APEX/e, $\Delta G_{\text{transfer}}$ has been reported to be -37.9 kcal/mol implying energetically allowed electron transfer. Of course, the PAG molecule must also be unstable with the electron addition for the electron transfer to lead to PAG decomposition.

The kinetics for photosensitized PAG decomposition is similar to direct excitation and appear first order to the PAG loading and exposure dose. The difference lies mainly in the dependence of photospeed on resist absorbance. For direct excitation mechanisms only photons absorbed by the PAG generate acid. Increased absorbance by the polymer is "wasted" and dramatically increases the dose required to image

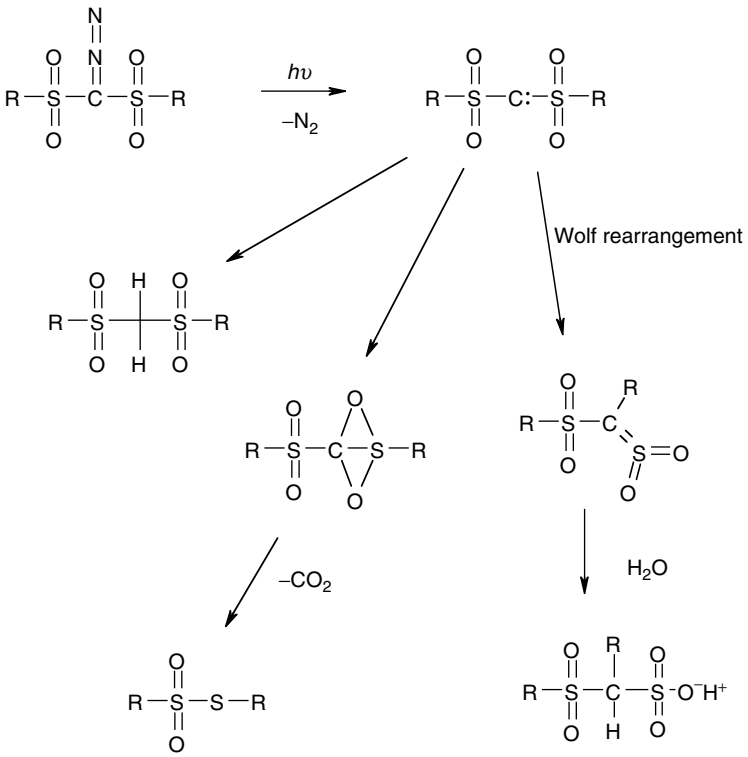


FIGURE 19.30 Reaction pathway for diazosulfonyl ester PAG decomposition.

the resist. If the polymer (or added donor compound) efficiently sensitizes the PAG decomposition, the increased resist absorbance only marginally affects its photospeed. The add absorbance is not wasted, but is channeled into useful photochemistry. This has been observed for dyed versions of DUV resists, where the absorbance of the resist has been increased to limit standing waves and increase process latitude. The observed dose-to-clear was only marginally affected by the increased absorbance from added dye compounds.

19.3.2.3 Post-Exposure Bake Step

After the exposure step produces a latent image of acid molecules in the resist, a thermal bake is implemented to drive the deprotection reaction to completion. This deprotection reaction is catalyzed by the acid generated during the exposure. It is this catalytic deprotection step that most distinguishes a chemically amplified resist, CA, from a conventional DNQ/novolac resist.

Many distinct protection group chemistries for CA resists have been developed. Only three of the protection chemistries, however, have found their way into widespread commercial acceptance: *t*-butyloxy carbonates, tertiary esters, and acetal/ketals. Each system has a unique set of advantages and disadvantages. Commercial resists systems, typically, contain a combination of protecting groups to achieve the desired properties. These resist systems may contain two or more protecting groups from the same class or may combine protecting groups from two separate classes. A common approach for 248-nm systems is to combine *t*-butyloxy carbonates and acetal protecting groups on a pHOST polymer backbone. When protecting groups from different classes are combined, the resist is classified as a hybrid resist system.

19.3.2.3.1 Post-Exposure Bake Chemistry

19.3.2.3.1.1 t-Butyloxy Carbonates

The original IBM *t*-BOC system, in which phenolic groups are protected by *t*-butyloxy carbonates, is still in widespread use today. As shown in Figure 19.31, the *t*-BOC protected pHOST deprotects under acidolysis to yield pHOST, isobutylene, and CO_2. The resulting phenolic sites render the resist soluble in an aqueous-base developer. The other photoproducts, CO_2 and isobutylene evolve out of the film as gasses. The increased amount of outgassing from chemically amplified resists may lead to photodeposition of contaminants on the last lens element. This aspect of chemically amplified resists must be considered, when the resist is designed and used in production to limit the amount of contamination.

The original IBM resist was conceived as a homopolymer with 100% protection of the phenolic sites. This presented several problems with processing that have subsequently been alleviated. First, the polymer is very hydrophobic and does not wet with aqueous developer requiring the use of an organic solvent developer. Second, the large volume of the protecting groups produces excessive shrinkage in the exposed areas. Furthermore, during plasma etch processing, the protecting groups readily come off reducing the etch resistance. These problems were addressed by formulating the resist using a partially protected polymer. Since, the solubility of the polymer, in aqueous developer, occurs only after 80% of the polymer is deprotected, a large percentage of the protecting groups are not needed in the functionality of the resist. Initial protecting levels vary, but are typically in the 25%–35% range. At this protection level the polymer wets readily with aqueous base developer and has only minimal unexposed develop rate. The shrinkage and plasma etch problems are also alleviated by only partially protecting the polymer, since a reduced fraction of the polymer becomes volatile. Several variants of the *t*-BOC protected scheme have been published, Figure 19.32,[83–85] yet partially protected pHOST remains the most predominant system using the *t*-BOC protecting group.

19.3.2.3.1.2 Tertiary Esters

An alternate chemistry based upon *t*-butyl esters developed by IBM researchers is also widely used for chemically amplified resists. Tertiary esters can be cleaved under acid catalysis at elevated temperatures to yield a carboxylic acid group. The carboxylic acid group is very soluble in aqueous base and this chemistry is very effective for resist systems. The deprotection mechanism is similar to that of *t*-BOC and is shown in Figure 19.33. Protenation of the ester group liberates a tertiary carbocation that undergoes β proton elimination to produce an acidic proton and volatile isobutylene.[86] The acid catalyzed

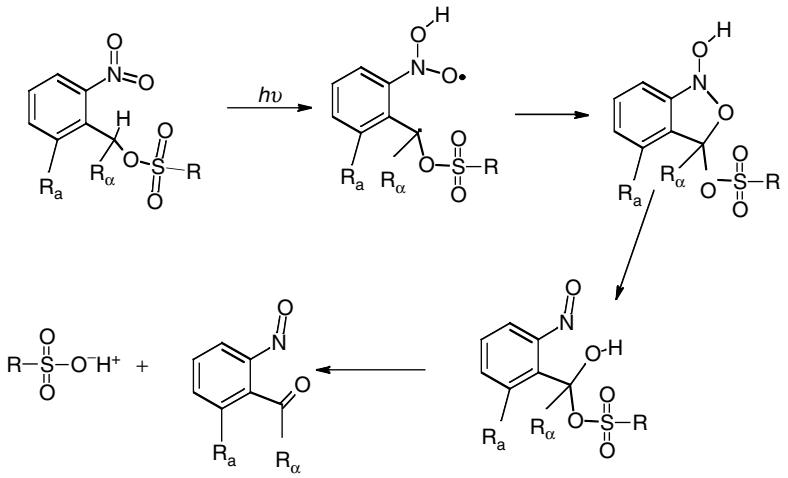


FIGURE 19.31 Reaction pathway for nitrobenzyl sulfonate ester PAG decomposition.

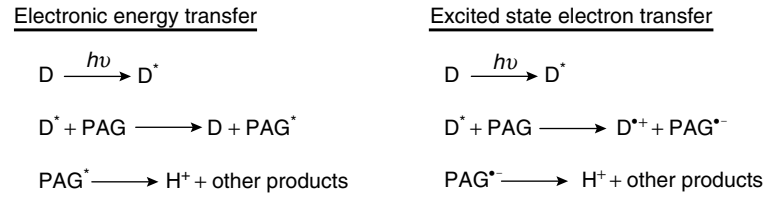

Electronic energy transfer
D —hυ→ D*
D* + PAG → D + PAG*
PAG* → H+ + other products
Excited state electron transfer
D —hυ→ D*
D* + PAG → D•+ + PAG•−
PAG•− → H+ + other products

FIGURE 19.32 Sensitization mechanisms for PAG decomposition.

deprotection is efficient at temperatures above 130°C, yet the uncatalyzed ester has a thermal stability above 180°C. A PEB above 130°C is, therefore, effective for delineating the exposed and unexposed areas of the resist. Other tertiary esters have been used as protecting groups as shown in Figure 19.34.[87,88] This class of materials is finding utility for chemically amplified resists at 193-nm exposure, where aromatic groups absorb too strongly precluding the use of phenol and *t*-BOC protected phenol. This will be covered in more detail in Section 19.3.2.3.1.3.

19.3.2.3.1.3 Acetals and Ketals

The use of acetal-blocked groups in acid catalyzed imaging systems was first proposed in 1973.[89] This invention predates the IBM *t*-BOC system, yet it received little or no attention until the late 1980s. The resist system designed by Smith et al. of 3M Company consisted of a tetrahydropyranol ether of novolac resin and a PAG. The tetrahydropyranyl group (THP) protected novolac is insoluble in a basic solution. Photogenerated acid catalyzes the hydrolysis of the polymer and produces a novolac resin that is soluble in an aqueous base. Researchers at Hitachi adapted this chemistry to 248-nm exposure by switching the novolac resin to pHOST with its improved optical properties.[90] The researchers at Hitachi studied the deprotection mechanism of this resist and proposed the chemistry shown in Figure 19.35.[91] This resist chemistry can be extended to other acetal and ketal protecting groups. Many experimental and commercial resist systems have been built upon this chemistry.[92–94] The general reaction is shown in Figure 19.36.

The protection/deprotection reaction for acetal/ketals is a reversible reaction that is driven to the deprotected state only when excess protecting group is driven away through volatization or secondary reactions, as with water is shown in Figure 19.35.

Weaker acids may be used to deprotect acetal/ketal systems than is needed for *t*-butyl ester or *t*-BOC systems. This characteristic of acetals has several advantages and disadvantages. At elevated temperatures the phenol group is acidic enough to catalyze acetal deprotection. Partially protected polymers will, therefore, autocatalytically deprotect, limiting the thermal stability of partially protected acetal resists. Higher levels of protection show higher thermal stability with the thermal decomposition temperature having a linear dependence on the level of acetal protection.

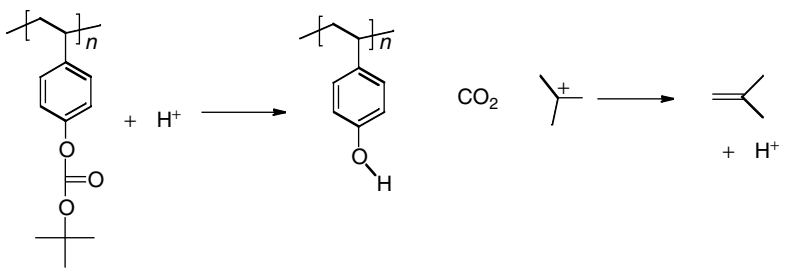


FIGURE 19.33 tert-butyloxy carbonates (*t*-BOC) deprotection mechanism.

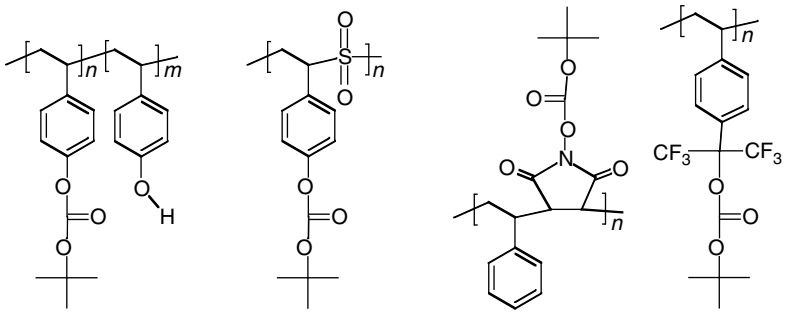


FIGURE 19.34 *t*-BOC protected resists.

Ketal protecting groups deprotect much more easily than acetals. This is attributed to the inherent stability of the tertiary carbocation intermediate formed during deprotection of ketals compared to the secondary carbocation formed during acetal deprotection. Ketal systems deprotect so rapidly at room temperature that a PEB is hardly necessary except to diffuse out standing waves.[95] This ease of cleaving makes ketal systems intriguing, but poses a serious problem for their implementation. The thermal stability in the presence of phenolic groups is so low for ketal systems that measurable deprotection may occur even at room temperature over a period of several weeks. The shelf life of these systems can be limited. Improvements in the thermal stablity and shelf life can be made by increasing the bulkiness of both acetal and ketal protecting groups and by introducing small amounts of base to the formulation.

19.3.2.3.1.4 Base Additives

Every commercial chemically amplified resist employs the use of one or more base additives. Initially the purpose of this base additive was to reduce susceptibility of CA resists to environmental contamination from airborne amines. The rationale being that a small amount of airborne contamination would affect only a minor change to the photospeed of the resist. However, it was later discovered that the introduction of base additives increased the working contrast of the resist,[96] reduced LER,[96] and facilitated the optimization of resist systems to specific feature geometries.[96] The imaging capability achievable using an acid, base, and polymer platform is much better than from a simple Acid and polymer platform. By choosing the right photoacid and base additive combination resists can be optimized for bright field (logic gates) or for dark field features (contact holes).

19.3.2.3.2 Post-Exposure Bake Kinetics

As with the exposure step, the PEB process can be described using a simple chemical mechanism. This mechanism is shown in Figure 19.37 and involves several concurrent reactions. First, the catalytic photoacid can react with polymer protecting sites, *M*; to generate deprotected sites, *X*; and to regenerate the catalytic acid. The photoacid can also react with base or quencher molecules, *Q*, within the resist to become an inactive species *A*–*Q*. This quenching reaction is an acid base equilibrium type reaction and is reversible. These quencher molecules can come from unwanted contamination or can be purposely added

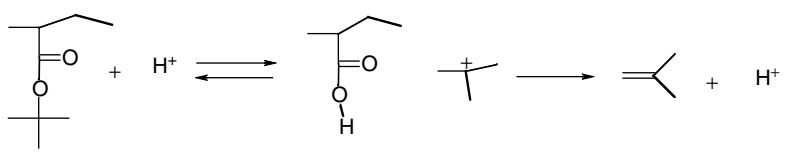


FIGURE 19.35 *t*-Butyl ester deprotection mechanism.

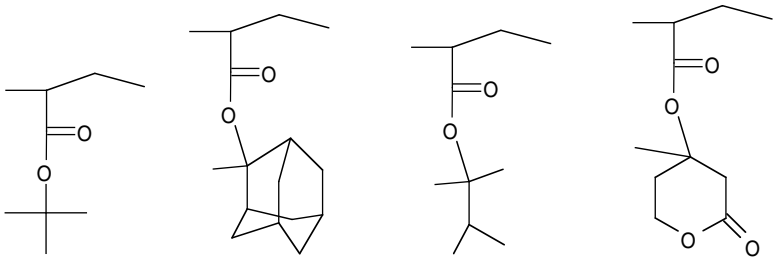


FIGURE 19.36 Tertiary ester protecting groups.

to the resist to alter its lithographic performance. The acid molecule can also undergo other reactions that render it ineffective for catalysis. This later loss reaction is typically unknown, but has been ascribed to both evaporation[97] and reaction with the polymer matrix.[97] A set of differential equation can be written to mathematically describe the time-dependent concentration of each species represented in this mechanism. These are given in Equation 19.7. Because the concentration of species is not constant throughout the film and mass transport is possible, diffusion terms, e.g., $D\nabla^2$, are required for each species.

$$\mathrm{M + Acid} \underset{\sigma D}{\overset{\sigma D}{\leftrightarrow}} \{\mathrm{M - Acid}\} \xrightarrow{k_a} \mathrm{X + Acid}$$

$$\mathrm{Q + Acid} \underset{k_{Q^-}}{\overset{k_{Q^+}}{\leftrightarrow}} \mathrm{AQ}$$

$$\mathrm{Acid} \xrightarrow{k_{loss}}$$

$$\frac{dM}{dt} = -\frac{\sigma D k_a}{\sigma D + k_a}[\mathrm{Acid}][M]$$

$$\frac{d[\mathrm{Acid}]}{dt} = -k_{loss}[\mathrm{Acid}] - k_{Q^+}[\mathrm{Acid}][Q] + k_{Q^-}[AQ] + D_A\nabla^2[\mathrm{Acid}] \tag{19.7a}$$

$$X(t) = M_{t=0} - M(t)$$

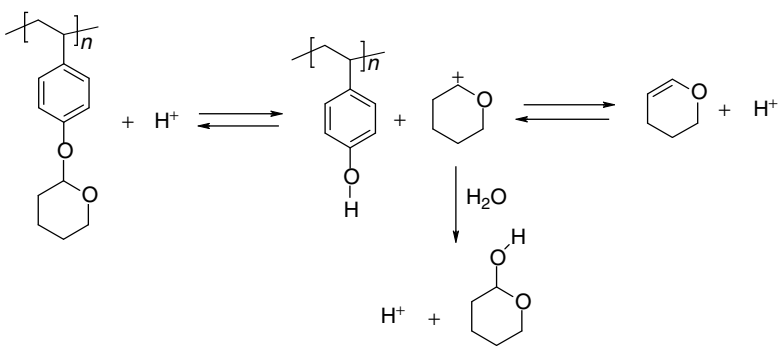


FIGURE 19.37 Tetrahydropyranyl group (THP) deprotection mechanism.

The bimolecular reaction between the protecting group and the acid is represented as a two-step reaction, where the two reactants must first come together and form a complex. This complex can either diffuse apart (the reverse reaction) or continue forward to produce the deprotected species and regenerate the catalytic acid. The complex formation and separation rates are diffusion limited and have the same rate constant, σD, where D is the diffusion coefficient of the acid molecule and σ is the cross-section number defining how close the two groups must be before the reaction can take place. The forward reaction is controlled by a localized reaction rate constant, k_a. The acid quenching reaction has both forward and reverse reaction rate constants. The ratio $K_{eq} = k_{Q^-}/k_{Q^+}$ is the acid-base equilibrium constant for this acid-base pair in the polymer matrix at the associated PEB temperature.

It should be noted that the mechanism shown in Figure 19.37 is just one possible description for CAR resists. A less-detailed mechanism can be used in which the bimolecular reaction is represented by just one overall rate constant. The less detailed mechanism for this step is

$$M + \text{Acid} \xrightarrow{k_a} X + \text{Acid} \tag{19.7b}$$

The rate equation for the protected group concentration is, then

$$\frac{\mathrm{d}[M]}{\mathrm{d}t} = -k_a[\text{Acid}][M] \tag{19.8}$$

where k_a is the overall rate constant for deprotection of the polymer. The functional form of the rate equation is the same for both mechanisms (Equation 19.7a and Equation 19.8), but Equation 19.7a distinguishes between diffusion limited and kinetic limited deprotection. If $\sigma D >> k_a$, then the overall rate reduces to k_a (kinetic limited). If $k_a >> \sigma D$, then the overall rate reduces to σD (diffusion limited). Classically, most bimolecular reactions in solid films are diffusion limited.

A similar contraction of the quenching reaction between the photoacid and the species Q can be made when the quenching is assumed to be instantaneous ($k_{Q^+} >> 1$) and complete (($k_{Q^-} = 0.0$)). Under this assumption, the amount of photoacid available for deprotection within each area of the photoresist is equal to the amount of photogenerated acid minus the localized amount of quencher molecule Q. This unquenched amount of photoacid is designated as $\text{Acid}_{\text{free}}$. The complete simplified model is shown in Figure 19.38. This is the standard model used to understand and simulate the PEB process for chemically amplified resists. The final rate law is given in Equation 19.9.

$$\frac{\mathrm{d}[M]}{\mathrm{d}t} = -k_a[\text{Acid}_{\text{free}}][M] \tag{19.9a}$$

$$\frac{\mathrm{d}[\text{Acid}]}{\mathrm{d}t} = -k_{\text{loss}}[\text{Acid}] + D\nabla^2[\text{Acid}] \tag{19.9b}$$

$$[\text{Acid}_{\text{free}}] = [\text{Acid}] - [Q] \tag{19.9c}$$

where M is the concentration of blocking sites on the polymer, X is the concentration of deblocked sites, Acid is the amount of photogenerated acid, Q is the amount of quencher, $\text{Acid}_{\text{free}}$ is the amount of unquenched acid, D is the acid diffusion coefficient, k_{loss} is an unspecified first-order loss process, and k_a is the overall deprotection rate constant.

The temperature dependence of the reaction rates and diffusion coefficient is modeled by an Arrhenius relationship.

$$k_a = A_a e^{-E_a/RT} \tag{19.12}$$

$$D = A_D e^{-E_D/RT} \tag{19.13}$$

where T is the PEB temperature in degrees Kelvin and R is the universal gas constant (0.0019876 kcal/(Kmol)). The activation energy (e.g., E_a) and pre-exponential factors (e.g., A_a) are determined

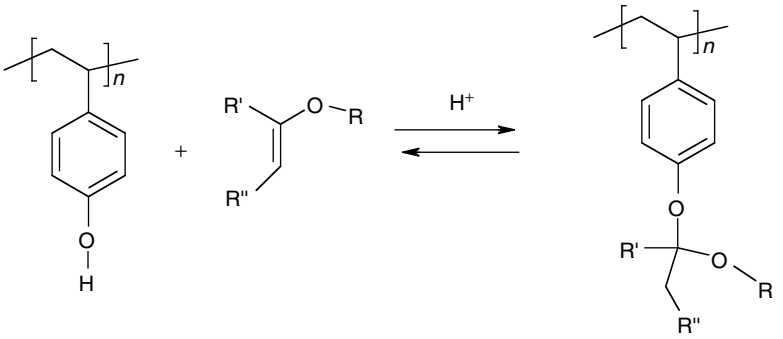


FIGURE 19.38 Acetal (R′ = H) and Ketal Protection of polyhydroxystyrene (pHOST).

experimentally. This temperature dependence is more appropriate for liquid or gaseous reactions. But in the light of the limited range of interest for each resist material, this functional form has proved adequate.

More complicated models for molecular diffusion in polymer films exist.[98] These are generally based upon the free volume theories proposed by Williams, Landau, Ferry,[98] and others.[99] The simple Arrhenius dependence of the diffusion coefficient is used in this work because of the difficulty in accurate measurements of diffusion through reactive films. Only a simple diffusion term can be measured with any confidence using techniques available today.

$$M + \text{Acid}_{\text{free}} \xrightarrow{k_a} \text{X} + \text{Acid}_{\text{free}}$$

$$Q + \text{Acid}_{\text{free}} \xrightarrow{\text{instantaneous}} \text{AQ}$$

19.3.2.4 Develop Step

The dissolution of a resist resin in aqueous developer has been described as a chemical reaction of the basic developer with the resin to deprotenate the polymer followed by rapid dissolution only if an appropriate number of polymer sites are deprotenated or ionized.[100,101] For conventional resists this ionization is inhibited by hydrogen bonding between the resin and PAC. Although effective, this inhibition through hydrogen bonding is not absolute. For CA resists, the potentially soluble site is protected from ionization by a covalent bond to the protecting group and the ionization is completely inhibited. The capability to covalently prevent this ionization of the resin permits chemically amplified resists to achieve develop rate contrasts that are much higher than conventional DNQ/novolac resists.

Unlike conventional DNQ photoresists, little published work has centered on fundamental models for the develop rate of chemically amplified resists. Initial work has focused on extending the phenomenological models for DNQ resists to DUV systems. These functional fits are generally successful, but give no insights into the chemical origins of the develop rate function for DUV resists.

The develop rate of a chemically amplified resist depends upon many factors including level of protection/deprotection, p*K*a of deprotected site, polarity of protecting group, molecular weight of polymer, developer strength, solvent content, PAG loading, presence of low molecular-weight dissolution inhibitors, thermal history, etc. To date a fundamental model has not been presented that can satisfactorily predict the influence of these various inputs apriori. However, several experimental papers have shed light on the various factors that have an effect on the development rate. It has been shown that the primary influences on develop rate for chemically amplified resists are the protecting/deprotected groups, molecular weight and developer strength. Other factors do play a minor, albiet, measurable role in develop rate. For instance, the develop rate has been shown to increase linearly with

solvent content.[102] Also various PAGs are effective as dissolution inhibitors.[103] These minor influences on the develop rate will not be discussed, further, here.

19.3.2.4.1 Protecting/Deprotected Groups

The develop rate as a function of the protection level (ratio of protected sites to the degree of polymerization) has been measured by many researchers for various resist systems.[104–107] Figure 19.39 shows this data for several platforms. For each resist system the data has a similar form. At high protection levels, the develop rate increases exponentially with decreasing protection levels. Eventually the develop rate levels off at some finite limiting value.

The slope of the log develop rate vs. deprotection level is a direct measure of the develop contrast for the resist. It is primarily a function of the p*K*a of the deprotected group, the polarity of the protecting group, and the developer strength. By increasing the acidity of the deprotected site, the develop rate increases. Hence, the inherent develop rate contrast is higher for carboxylic acid (p*K*a ~ 4) compared to phenolic (p*K*a ~ 10) based chemically amplified resists.

Iwasa et al. studied the influence of the protective group polarity on the dissolution rate of partially protected polymers. In this study, the polarity of each protective group was represented by the relative dielectric constant of its model compound. The dissolution rate was found to increase exponentially with the relative dielectric constant of the protective group.

Because any completely protected polymer has an extremely low develop rate it is quite common for resists to be formulated with only a partially protected resin. This allows tailoring of the unexposed resist develop rate, photospeed, wettability, and adhesion. As shown in Figure 19.39, phenolic resins do not show significant develop rate until approximately 80% of the polymer sites are deprotected. It is, therefore, advantageous to formulate a phenolic-based chemically amplified resist with only 20%–40% of the phenolic sites protected. The unexposed resist will have just enough solubility to enhance wetting by

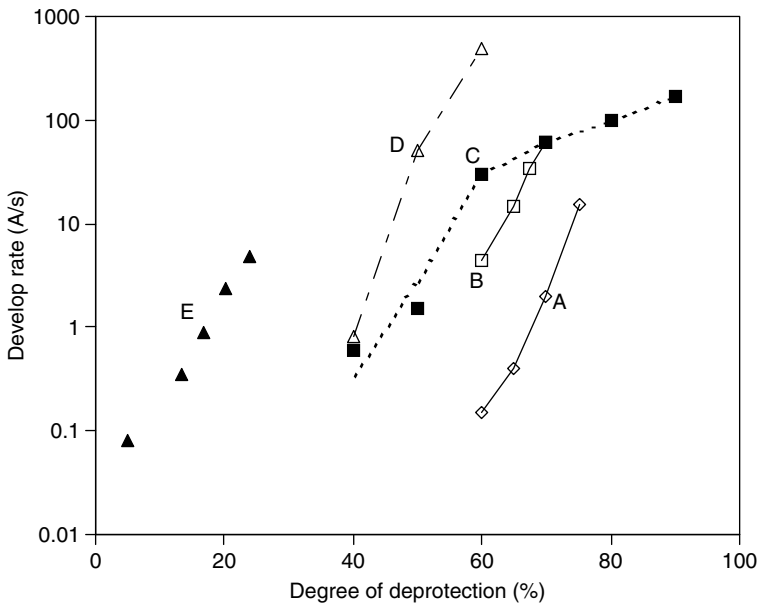


FIGURE 19.39 Dissolution rate of various chemically amplified resist resins as a function of deprotection. (Data from (A) Itani, T., et al. *J. Photopolym. Sci. Technol.* 10 (1997): 409.; (B) Ito, H., and E. Flores. *J. Electrochem. Soc.* 135 (1988): 2322; (C) Itani, T., et al. *Proc. SPIE* 2438 (1995): 191; (D) Iwasa, S., et al. *Proc. SPIE* 3049 (1997): 126; (E) Yamachika, M., et al. *J. Photopolym. Sci. Technol.* (1999): 12.)

the developer and adhesion to the substrate. Also, the PEB step will be more controllable as the number of catalytic events needed to produce imaging is minimized.

For carboxylic acid-protected polymers the develop rate becomes significant at much lower levels of deprotection (ca 20%). Because the develop rate contrast is high for carboxylic acid groups, it is a difficult task to balance the wettability and adhesion of these polymers by partially deprotecting the polymer. When enough free carboxylic acid is incorporated into the polymer to achieve sufficient wetting, the unexpose develop rate becomes so high that the apparent contrast of the resist is decreased.[108] Other non-acidic but polar groups must be incorporated into the polymer, if carboxylic acid groups are the sole solubility switch. For 248-nm systems the most successful approach has been that used in the ester capped (ESCAP) polymer, where an unprotected hydroxystyrene is copolymerized with *t*-butyl ester protected carboxylic acid groups. The unprotected phenolic groups are incorporated at significant levels to achieve good adhesion and wettability without compromising unexposed develop rate and contrast. The uses of acid labile protected carboxylic acid groups permit high develop rate contrast. Further, improvements of the resist performance are possible by incorporating non-soluble, yet etch resistant monomers, into the polymer. The relative loading of soluble, etch resistant, and acid labile groups is optimized to achieve the desired develop rate characteristics for maximum process lattitude for individual feature types.[109]

19.3.2.4.2 Molecular Weight

The molecular weight of the chemically amplified resist resin plays a secondary role in the slope of the develop rate vs. deprotection level curve of Figure 19.40. Molecular weight does, however, affect both the intercept of the develop rate curve and the leveling off point for the maximum develop rate.

As with novolac resins the develop rate slows with increasing molecular weight. The maximum develop rate for equivalent molecular-weight polymers of pHOST, however, are approximately 10 times faster than novolac polymers. The increase of intramolecular hydrogren bonding in pHOST, over intermolecular bonding in novolacs, is given as the explanation of this phenomenon.[110] As a

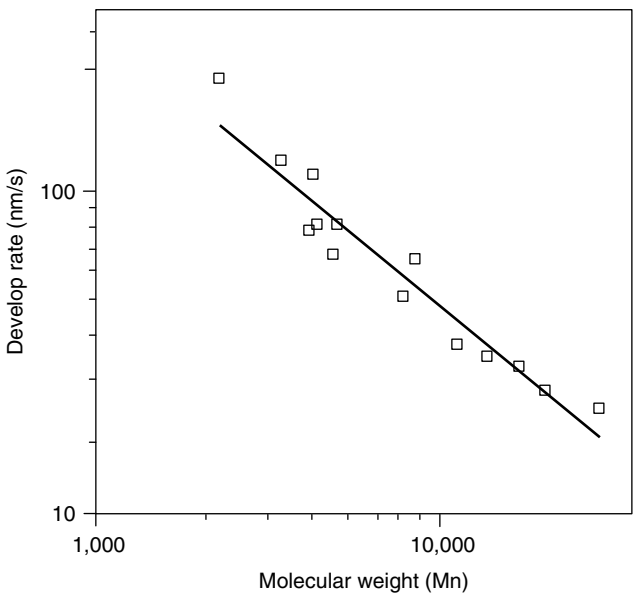


FIGURE 19.40 Dissolution rate of pHOST polymers of various polydispersity as a function of number averaged molecular weight (Mn).

consequence, the polymer molecular weights used for DUV resist systems has generally been higher than for novolac resins. Weight averaged molecular-weight values between 3000 and 10,000 are common.

To determine the role of molecular weight and polydispersity on the maximum develop rate of pHOST, researchers at Shipley synthesized very narrow dispersity polymers using "living" free radical polymerization.[111] The narrow well-defined molecular-weight resins were, then blended together to obtain multiple molecular-weight mixtures of known polydispersity. They found a very strong correlation of the develop rate to the number averaged molecular weight, Mn, regardless of the polydispersity (Figure 19.40). Because of the extremely high develop rate for carboxylic acid polymers, no detailed study of maximum develop rate vs. molecular weight has been published for these polymers.

19.4 ArF Materials, Immersion Lithography and Extension of ArF

19.4.1 ArF Materials

Although the resolution limit of 248-nm resists has continued downward through improved resist design, increased Lens NA, and other optical enhancements, it was recognized early in the life cycle of 248-nm resist that shorter wavelengths would eventually be required. The continued quest of shorter wavelengths has led to the development of resist systems that work at 193 nm and shorter wavelengths. One ninety three-nanometer exposure systems rely upon an argon fluoride (ArF) excimer laser source. The basic resist relief image formation mechanism employed for 193-nm systems is functionally equivalent to the acid catalyzed deprotection mechanism used for 248-nm systems. The chemically amplified resist concept, first described two decades past and originally targeted for the 1000-nm device generation, has proved to have remarkable versatility. The semiconductor industry has come to rely on the properties of chemically amplified resists to achieve high resolution, high aspect ratio imaging accompanied by the high throughput that stems from their catalytic imaging mechanism. As the industry maps the evolution of lithographic technology to the 32-nm regime, it is appropriate to review the factors that control the performance of chemically amplified, and examine whether the traditional evolutionary path of materials refinement will provide materials capable of supporting device manufacturing at those dimensions. The impacts of image blur, line edge roughness (LER), and shot noise on the ability to image chemically amplified resists at nanoscale dimensions need to be understood. The rapid progress that has characterized the semiconductor industry, since its birth in large part stems from refinement of the lithographic techniques used to fabricate integrated circuits. Industry planning calls for the pace of miniaturization in semiconductor technology to be maintained well into the future.[112] The prospects and issues tied to the extension of semiconductor technology into the nanoscale regime have been examined in detail,[113] and factors that limit the use of lithographic exposure technology to support that miniaturization have been recently reviewed.[114,115]

Embedded in the specifications of the industry roadmap is the need for chemically amplified resists that provide lithographic performance suitable to sustain their extension to the 32-nm dimensional regime. It is recognized that the advancement of semiconductor technology cannot continue at the current pace. Given the economic importance of semiconductors, the nature and positioning of various limits has been examined in some depth. Such organization has facilitated a systematic evaluation of the theoretical and practical factors that will influence the evolution of semiconductor technology. The limits of lithography can be considered in a similar framework. The ultimate achievable resolution, radiation sensitivity, and preciseness of image formation are the consequence of a set of fundamental limits controlled by many factors. These set a lower bound for the next hierarchical level of material limits, which are shaped by the intrinsic chemical and physical properties of the imaging medium (the resist) and may be degraded from those at the fundamental level. The material limits, in turn, form a lower bound for process limits, where the attributes of the tooling and operating conditions used in the imaging process dictate the best achievable lithographic performance. For example, the wavelength and

NA of the exposure tool may restrict overall resolution to a level that is inferior to the intrinsic resolution of the resist in use.

19.4.2 ArF Transparent Polymer Systems

The need for new polymer systems is solely based on the optical properties of existing KrF materials. Figure 19.41 demonstrates the highly absorbing nature of phenolic systems at 193 nm. The first single layer transparent polymer system discussed by Allen et al. was a single layer acrylate system originally designed for printed circuit board applications over 15 years ago.[116] Improvements in etch resistance quickly became a priority by incorporating cyclic olefins into the backbone.[117] Over the past 10 years, numerous researchers have developed new materials or improvements to the existing materials to enhance etch resistance and imaging performance. The literature has numerous examples; however, for the purposes of this book chapter we will focus on several basic systems. The chemical structures of the polymers are shown in Figure 19.42a and b. These cyclic olefin polymers[118] are thermally stable, another class consists of poly (2-methyl-2-adamantyl methacrylate$_{50}$ − 2,6-norbornencarbolactone methacrylate$_{50}$) as a resin. The second consists of polyacrylate. The third consists of poly (*t*-butyl-cycloolefin$_{50}$-maleic anhydride$_{50}$) (COMA) shown in Figure 19.43a–c[119] along with number combinations of all the above systems.

19.4.2.1 Typical ArF Polymer Systems

Figure 19.44 is typical 193-nm polymer systems[117] discussed early by researcher at the University of Texas, IBM, and others to develop systems that created high quality image along with adequate etch resistance.

In 193-nm resist systems, neither the excited state nor electron transfer mechanisms are allowed energetically. For these systems the direct excitation pathway is the dominate mechanism for photoacid generation. Because small fractions of light absorbed by the polymer system generates the desired photochemistry, significant effort for polarity change.

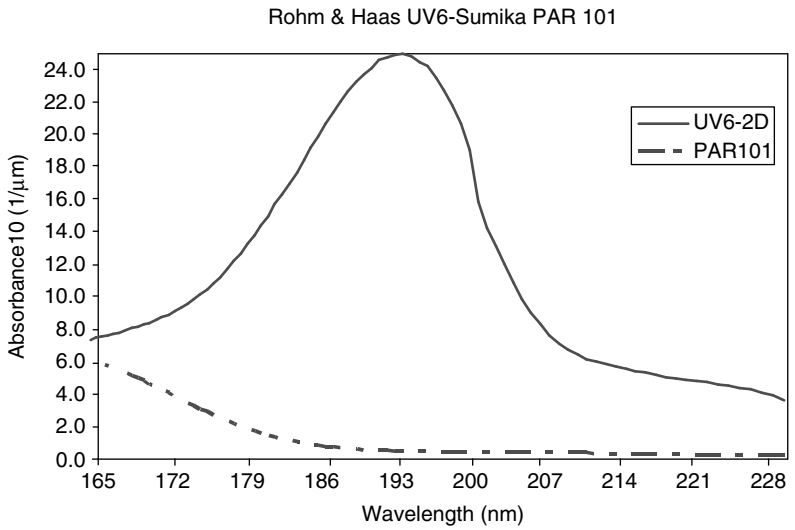


FIGURE 19.41 UV of KrF resist at 193 nm.

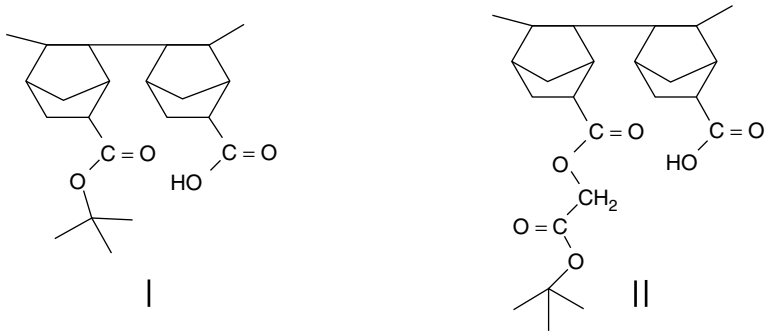


FIGURE 19.42 Polymer 1 is a ArF cyclic olefin copolymer. Polymer 2 has slightly lower thermal stability due to CH_2 spacer group.

19.4.2.2 Exposure Step

In 193-nm resist systems, neither the excited state nor electron transfer mechanisms are active. The oxidation potentials for polymer 193-nm resist polymer systems are too high to support electron transfer sensitization of PAG molecules. For these systems the direct excitation pathway is the dominate mechanism for photoacid generation. Because none of the light absorbed by the polymer system generates the desired photochemistry, significant effort has been made to produce more transparent polymer systems and higher yielding PAGs.[97]

For the direct excitation model the observed Dill *C* rate constant (Equation 19.10) can be related to molecular quantities.

$$C = \frac{\varepsilon\phi}{N_A}\frac{\lambda}{hc}2.303 \times 10^3 \tag{19.10}$$

where ε is the molar absorbance of the PAG and ϕ is the quantum yield for the photoreaction. Several researchers have measured the quantum yield for acid generation using different PAG's and polymer systems at both 248- and 193-nm exposure wavelengths. As seen, the range of *C* parameters observed in CA resist systems is generally within the 0.01–0.08 cm^2/mJ range. For typical commercial resists the exposure energy required to effect significant solubility switch is approximately 10 mJ/cm^2. This translates into 10%–50% of the initial PAG being decomposed into catalytic acid at working conditions for CA resist systems.

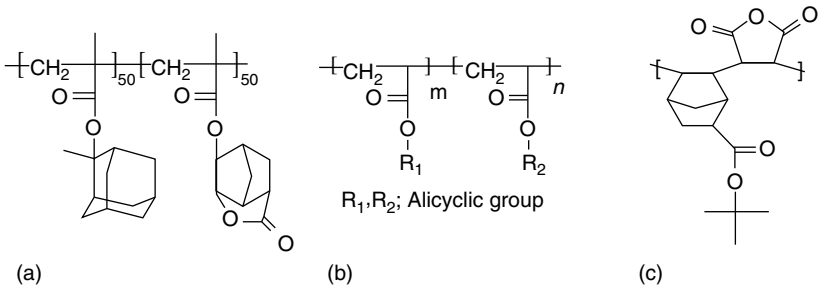


FIGURE 19.43 Example of polyacrylate systems.

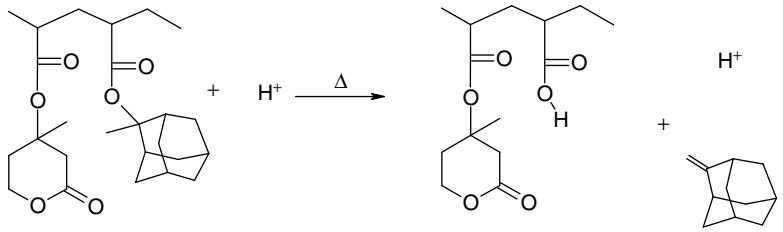


FIGURE 19.44 A typical acid catalyzed thermolysis reaction of 193-nm resists systems showing the generation of free carboxylic acid from a tertiary ester.

19.4.2.3 Deprotection Kinetics

After the exposure step produces a latent image of acid molecules in the resist, a thermal bake is implemented. During this thermal bake step, a thermolysis reaction is catalyzed by the photogenerated acid. For 193-nm resist systems, the thermolysis reaction typically generates a carboxylic acid from a tertiary ester as shown in Figure 19.44. Several events occur during the PEB and successful simulation of the chemically amplified resist process depends critically upon the correct PEB model and parameters.

19.4.2.4 Line Edge Roughness

Line edge roughness, for the purposes of this discussion,[120] is defined as the root-mean-square (rms) deviation of a single edge from a straight line. Its existence is one of the critical issues for the realization of sub-70-nm lithography. The subject alone could be the subject of an entire chapter in this book; however, we will only take a brief look into some of the cause and provide some examples. Line edge roughness deteriorates the accuracy and repeatability in measuring CD and the roughness is transcribed into the etched pattern. It has an effect on device performance, such as leakage, current, and threshold voltage. Engineers have studied its effects on device performance,[120,121] major cause,[122] and how to control it.[123–125] Photoresist researchers have studied several factors having an effect on LER in terms of resist materials along with variations in process or additive materials to reduce LER. Over the past few years the literature has numerous examples and data on various contributors to LER from a photoresist perspective. They include molecular weight and polydispersity of polymer; the volatility, diffusivity, and transparency of PAG; the basicity, volatility, and diffusivity of the quencher; polymer swelling, activation energy, and size of the protecting group and polymer type. Figure 19.45 are 100 nm l/s generated with a COMA and acrylate-based polymer system demonstrating the difference in LER just from materials.

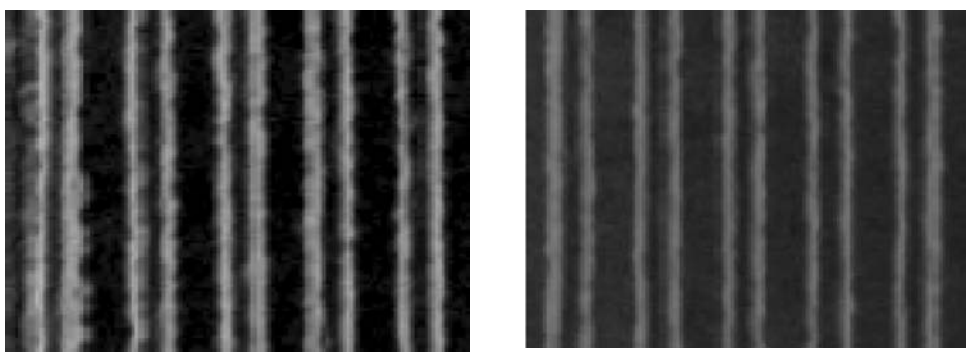


FIGURE 19.45 100 nm l/s demonstrating line edge roughness (LER).

19.4.2.5 New Classes of ArF Polymers

The development of polymers to meet the transparency requirements for 157-nm lithography created a class of polymers that consisted of high contents of fluorine. These systems[126] incorporated nearly 50% fluorine to achieve transparency goals. The unfortunate demise of 157-nm technology did create a vast library of knowledge in new systems that not only have nearly 99% transmission at 193 nm, but also had unique properties that improved the performance of various types of polymers. In this section, the authors will discuss these new classes of polymers along with new systems that have been created as topcoat or protective layer materials for immersion lithography.

The Willson Research Group at the University of Texas[127] explored the selective incorporation of fluorine in a norbornane system. The plot in Figure 19.46 demonstrates the improvement in absorbance at 157 nm of norbornane dependant on the location of the flouro group. In this plot we also see a significant improvement in absorbance at longer wavelengths.

This activity yielded several interesting polymers with low absorbance initially at 157 nm and later at 193 nm. The polymer shown in Figure 19.47 is copolymer of NBHFA and NBHFA *t*-BOC. Trinque et al.[128] discuss the synthesis and application of this polymer for imaging at 157 nm. Further, investigation into the optical properties of this system and imaging capability has also been investigated[129] that this copolymer is 99% transmissive at 193 nm.

Recently, Varanasi et al.[130] published variations of polymers shown in Figure 19.48, which takes advantage of simple free radical polymerization of acrylate systems that have incorporated norbornane for etch resistance. Up to now, we have discussed the incorporation of fluorine for improvements in transparency, which is still true, however, in this work not only is there an improvement, but Varanasi et al. discovered that the incorporation of a monomer containing fluorine assists in reducing swelling in acrylate polymer systems during development. Varanasi reported that since the p*K*a of HFA is similar to that of phenol, that HFA incorporated methacrylate resists would behave similar to ESCAP-based KrF resists in terms of resist dissolution kinetics. For the purpose of a comparison study, Varanasi prepared a simple copolymer of *t*-butylmethacrylate and NB-HFA-MA (40/60) using free radical polymerization method. This composition was chosen, primarily, to mimic well-known ESCAP copolymer of *t*-butylacrylate and *p*-hydroxystyrene (40/60). The corresponding resist formulation was prepared using industry standard PAG and quencher combinations. Dissolution rate vs. exposure dose curves were obtained by flood exposing (254-nm wavelength, obtained from Hg–Xe lamp) cast resist films at various exposures doses, processed and, then obtained dissolution rate information using quartz-crystal microbalance (QCM) method. The comparison of data shown in Figure 19.49 reveals that HFA-based ArF methacrylate resist behaves similar to ESCAP KrF resist, and do not show any swelling behavior even at the onset of dissolution contrast, unlike typical ArF methacrylate resists.

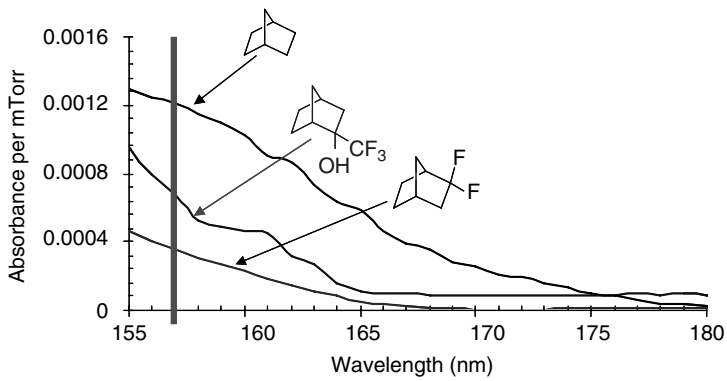


FIGURE 19.46 Absorbance data of norbornane and fluoronorbornane derivatives.

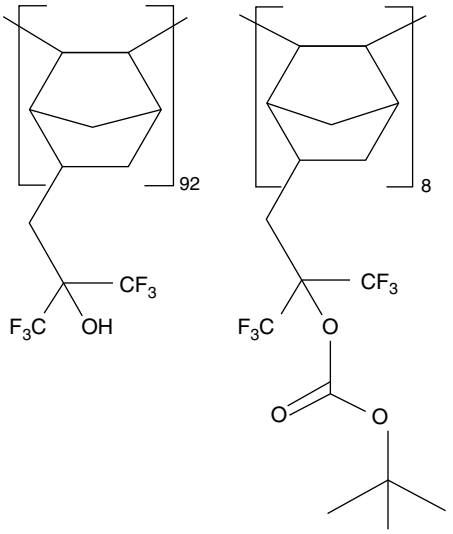


FIGURE 19.47 Fluoropolymer of NBHFA and NBHF *t*-boc.

Another interesting aspect of these systems is the improvement in PEB sensitivity. Typically high etch resistant methyl acrylates resists are based on multi-cyclic bulky protecting groups such as methyl adamantyl group. Resists derived from methyl adamantyl protecting group-based polymers often suffer from higher PEB sensitivity (5–10 nm/°C) with these systems reporting PEB sensitivities approximately 1 nm/°C.

19.4.3 Extending ArF

ArF immersion lithography has emerged as a promising candidate for 65-nm node technology.[131] The basic idea of immersion lithography is filling the gap between the final lens element and the photoresist with a fluid, which has a higher refractive index (n) than air ($n=1$) so that resolution and (depth of focus) DOF can be increased.[132] Figure 19.50 depicts the two advantages of immersion technology. One is

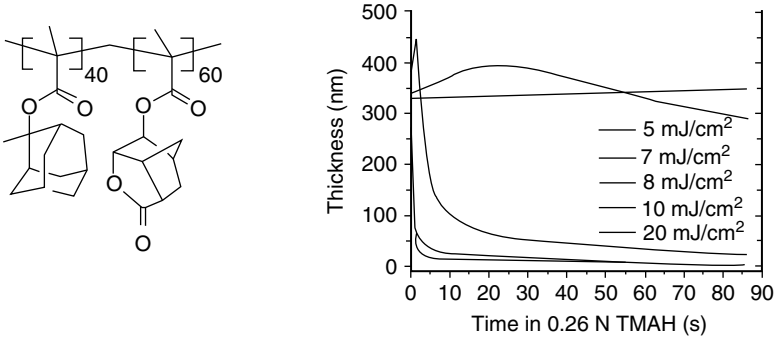


FIGURE 19.48 Dissolution rate vs. exposure dose curves generated for state-of-the-art KrF ester capped (ESCAP) and ArF (Methacrylate Resists).

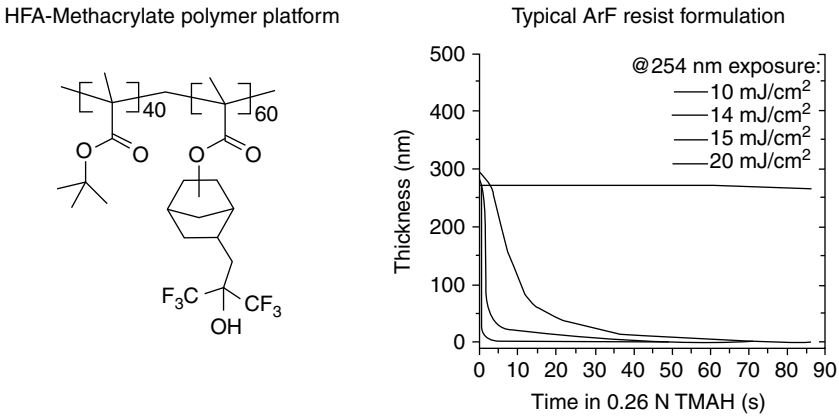


FIGURE 19.49 Dissolution rate vs. exposure dose curve obtained with an ArF resist formulation containing copolymer of *t*-butylmethacrylate and NB-HFA-MA.

to increase DOF of an exposure system, while maintaining same resolution of a dry system at equal NA. The image-forming angle of the deflected light in the photoresist does not change, but the incident angle in the fluid above the resist surface does change. Because the incident angle in the fluid becomes smaller, the available DOF is increased. Existing dry scanner lenses need little modification on the shape and position of the lens elements to preserve the incident angle in the resist. For NA beyond one, the advantage is to enhance the resolution beyond the limit of a dry system using the same vacuum wavelength. The optical system is re-designed to preserve the physical angle in the coupling medium. The incident angle of the exposure light in the resist can then be enlarged to resolve features in smaller half

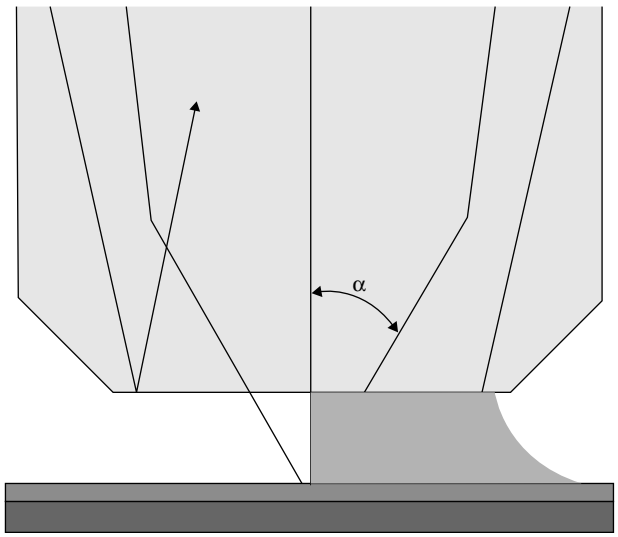


FIGURE 19.50 The two advantages of immersion lithographic system, (a) increase depth of focus by decreasing the incident angle in water, and (b) enhance resolution by enabling hyper NA lens design.

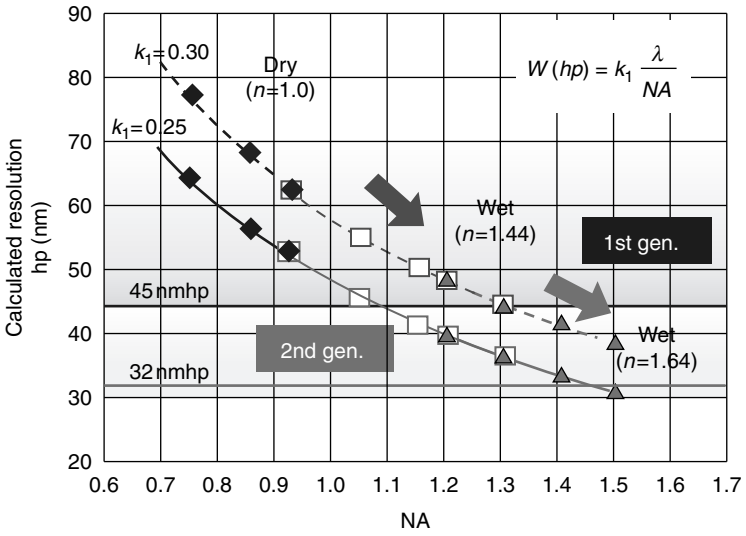


FIGURE 19.51 Calculated resolution vs. NA.

pitch. Of course, the incident angle does not have to be confined to only these two specific cases. Resolution and DOF can be traded off against each other by selecting the incident angle properly.

The success of ArF water immersion lithography is inspiring many engineers and scientists to think, if ArF immersion lithography could be put forward further. Could CD smaller than 45 nm, for example, 32 nm, be achieved by ArF immersion lithography with a high refractive index fluid? Figure 19.51 shows the calculated resolution (W) based on the Rayleigh equation; (Equation 19.1) where k_1 is process constant and is related to the difficulty of lithography process and has the lower theoretical limit of 0.25, λ is the wavelength, and NA is numerical aperture of the optical system.

19.4.4 Topcoats for Immersion Lithography

During the initial introduction of water immersion lithography, photoresist companies quickly discovered that existing ArF photoresists produced reasonably good lithography. SEMATECH sponsored an Immersion Task Force, which quickly investigated a number of aspects of photoresist chemistry. A series of surface experiments were performed ranging from contact angle, to investigate any surface energy changes, to XPS and TOF-SIMS to understand the contents of the film.[133] These investigations quickly pointed the industry in the direction to understand the surface interactions and components from the photoresist that leach into the water. These studies investigated the use of model resist systems based on copolymer of methyl-adamantyl methacyralate and γ-butryl lactone methacyralate along with three commons PAGs shown in Figure 19.52.[134] Data presented in Figure 19.53 was a clear indication on the amount of PAG that was leaching from the resist surface, however, there was a surprise that the perfluoro-octanoic sulfonic acid (PFOS) system had higher concentrations of PAG in the water and that the triflate system was less.

TOK developed a "cover material" called TSP-3A, which was a fluoropolymer that was cast over the ArF photoresist. The purpose of this cover material was to prevent any leaching and improvements in image quality. This material was insoluble in developer and required a separate solvent for removal. Due to the high fluorine content of the polymer the contact angle was extremely high, which lead to a number of other problems.[135] The industry quickly developed "top coats" that are developer soluble with lower

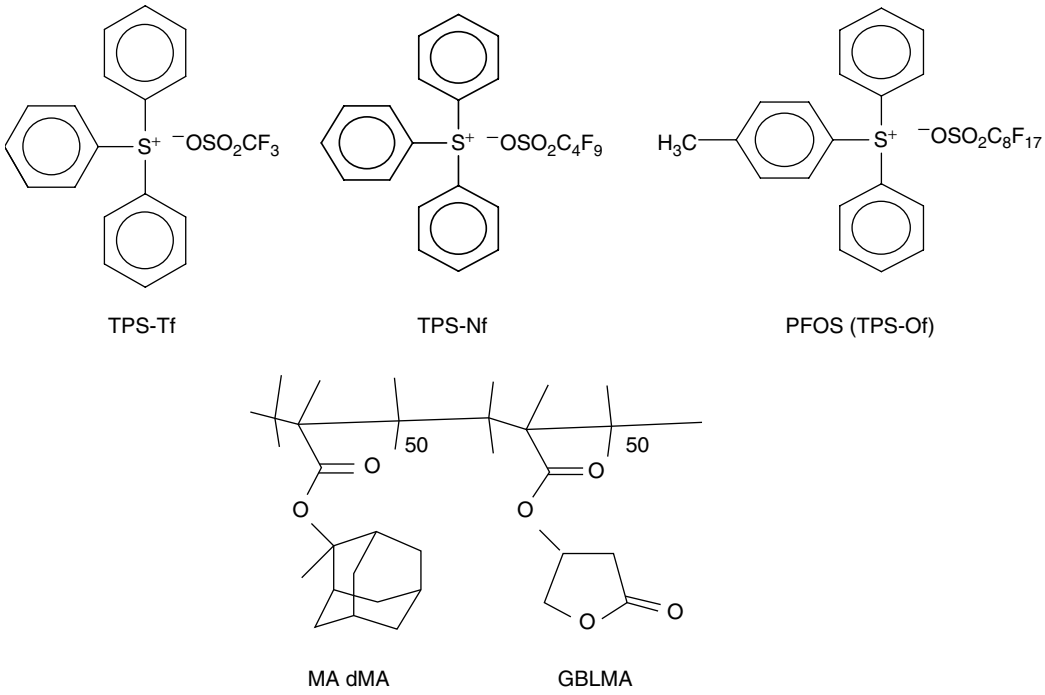


FIGURE 19.52 Photoacid generators and polymer for leaching studies.

contact angle.[126] This ability to quickly develop these systems is a benefit of the vast amount of material that occurred during the 157-nm development programs. As previously discussed, the highly fluorinated materials were used to gain the necessary transparency needed at 157 nm and the benefit was virtually 99.5% transmissive materials. These cover coats are excellent in the reduction of leaching, but not the total prevention.[136]

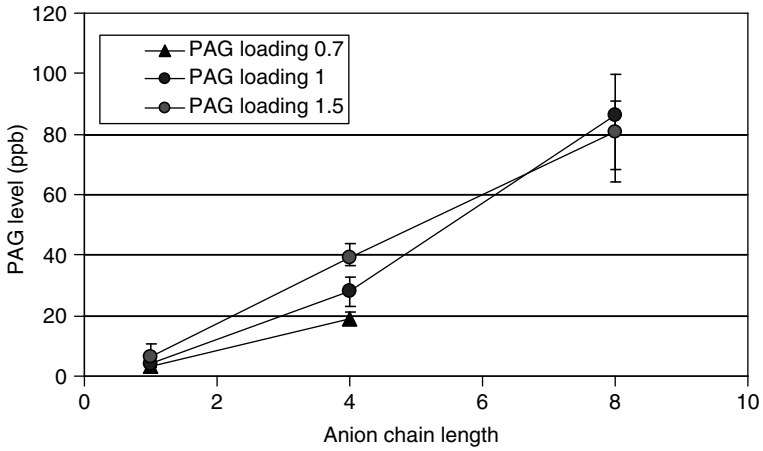


FIGURE 19.53 PAG leaching vs. anion chain length.

19.4.5 New Immersion Fluids

The further extension of ArF immersion can, in principle, continue if a fluid exists with physical properties similar to water, but maintains a higher refractive index at 193 nm. This increase in refractive index allows lens designers to build a larger lens system of greater NA and thus higher resolution.[137,138]

Water, as an immersion fluid, has a theoretical limit in NA equal to the index of water.[133] The practical limit for lens design is even less and estimated to be approximately 1.3 NA. With k_1 of 0.27, this would result in 40-nm half pitch resolution. The latest experimental data on high index fluids is presented in the paper of Sewell.[127] Burnett[125] pointed out that next to high index fluids also high index glass materials are required to enable the super high NA lens designs. Regarding lens designs, immersion lenses may follow two different approaches. The first one is the approach with a flat surface near the image side, the second one with a curved surface near the imaging side. With the flat surface approach, the refractive power is dominated by the glass material and the fluid index should be matched as good as possible to the index of the glass. The advantage of this approach is that the fluid film can be relatively thin. This relaxes the absorption requirements on the fluid. With the approach of a curved last lens surface, only the fluid index determines the maximum NA. However, in this case, the optical path through the fluid cannot be small, and thus, the requirements on the fluid absorption become very tight. Besides absorption, there are additional requirements on the fluid, like viscosity, thermal dependency, and cost. If we compare the basic requirements with the published experimental data[123–125,127] we conclude that the current fluids are too high in absorption, too high in dn/dT, and too expensive. If we assume the condition $n = n(\text{fluid}) = n(\text{glass})$ and assume maximum NA $= 0.9n$ and minimum $k_1 = 0.27$, we can plot the resolution limit of ArF immersion lithography. The result is shown in Figure 19.54. With the currently published index number of fluids and glass materials, ArF resolution is limited to 36 nm. In order to reach 32 nm, new fluid and glass materials are required with refractive index numbers exceeding 1.8.

From this calculation, 32 nm or below resolution can be achieved with high refractive index fluid ($n = 1.64$). Although extreme-ultraviolet (EUV) (13 nm) lithography has been suggested to be used in 32-nm node or below, the development of exposure tools for EUV is still in early stage and much time and effort is thought to be needed because of the technical hurdle. By making use of existing water

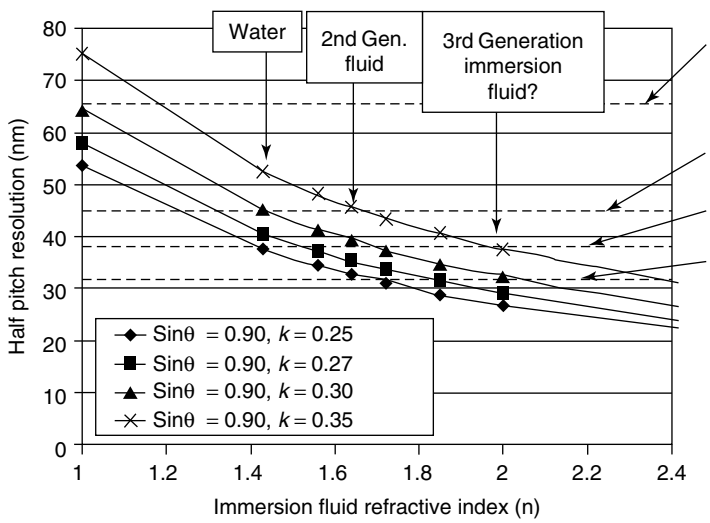


FIGURE 19.54 Resolution vs. immersion fluid refractive index.

immersion technology, ArF immersion with high index fluids has apparently the advantage of lower cost and risk for tool development. This is why ArF immersion is now gaining more and more spotlight as a candidate for the next generation lithography (NGL) technology. Initial attempts to develop high refractive index (RI) fluids for ArF immersion has been carried out aiming at increasing refractive index by addition of inorganic materials. Smith et al. reported various refractive indexes at 193-nm wavelength with doped waters. They utilized "charge-transfer-to-solvent" (CTTS) transition to induce the small absorption near the 193-nm wavelength with inorganic ions, and therefore, heightened the refractive index of water. They presented the result of 68 nm L–S imaging by an aqueous solution of 85% phosphoric acid with refractive index of 1.55 at 193 nm wavelength.[133] A unique approach is also reported by applying nano-sized metal oxide. Researchers at SEMATECH and Clemson University reported that refractive index of water dispersed with aluminum oxide nano-particles could be as high as 1.6.[134] Although this kind of an approach can take advantage of some favorable properties of water, they appear to sacrifice others. For example, although CTTS can increase the refractive index of water, it also reduces the transmittance of water. Inorganic ions of metal oxides can damage lens and or leave photoresist defects. Furthermore, mixed aqueous compositions have another disadvantage, the difficulty to precisely control the accuracy of their refractive indexes, as small amounts of variation in concentration would cause enough fluctuation in refractive index. The ideal solution would be a single component fluid.[139] Recently, researchers from JSR and Dupont disclosed organic fluids with a refractive index of 1.65 at 193 nm. Imaging studies have been completed through the use of interferometric lithography demonstrating 32 nm 1/2 pitch imaging. This demonstration is a great step forward in the further extension of immersion ArF lithography; however, there are still numerous challenges not only in fluids, but resist materials and the optical system of the exposure tool.

19.4.6 High Refractive Index (RI) Polymers

The idea of increasing the refractive index is not a relativity new concept; however, understanding the impact is.[139] Recent studies at SEMATECH and the University of Queensland[140] have focused on the incorporation of sulfur into the polymer. The results have demonstrated increases in refractive with relatively small amounts of sulfur incorporation. Presently the vast majority of ArF polymers have a refractive index of approximately 1.7. Figure 19.55 is the structure of a typical ArF acrylate polymer system.[141] Figure 19.56 is the structure of a sulfur containing copolymer,[142] and Figure 19.57 is a UV spectrum of each polymer demonstration, the increase in refractive index. But, why increase the refractive index? Figure 19.58 is a plot of exposure latitude vs. refractive index. This plot demonstrates with increase refractive index improvements in exposure latitude can be achieved. The theory has been previously discussed[143] and Figure 19.59 is the individual process capability plots for polymers with

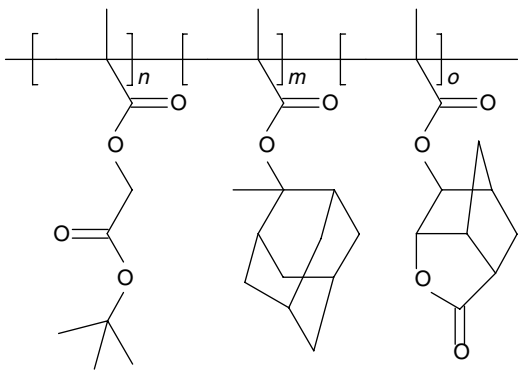


FIGURE 19.55 Structure of standard ArF polymer.

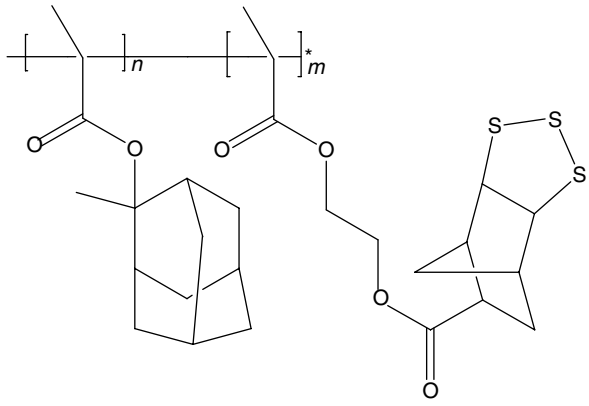


FIGURE 19.56 Structure of sulfur containing copolymer.

increasing refractive from 1.69 (current polymers) up to 2.29 demonstrating the increased exposure latitude for a 50 nm-image on a 130-nm pitch using water as the immersion fluid and an NA of 1.35.

19.4.7 Post-ArF-Material Requirements

At this time (mid-2005), 90-nm device fabrication is continuing to ramp up. The International Technical Roadmap for Semiconductors[112] (ITRS), which outlines target device and materials requirements for future generations of semiconductor devices, calls for device dimensions to shrink to approximately 20 nm minimum size by the year 2016. It is anticipated that the NGL exposure technologies[115] using EUV[144] radiation or electron beam projection[145] (EBP) will be necessary to achieve adequate resolution. It is not surprising that resist functional requirements become increasingly stringent as dimensions of the target devices shrink. For the ITRS 22-nm technology node (dynamic random access memory half-pitch), which is the most stringent metric for resist resolution rather than the less reliable measurement

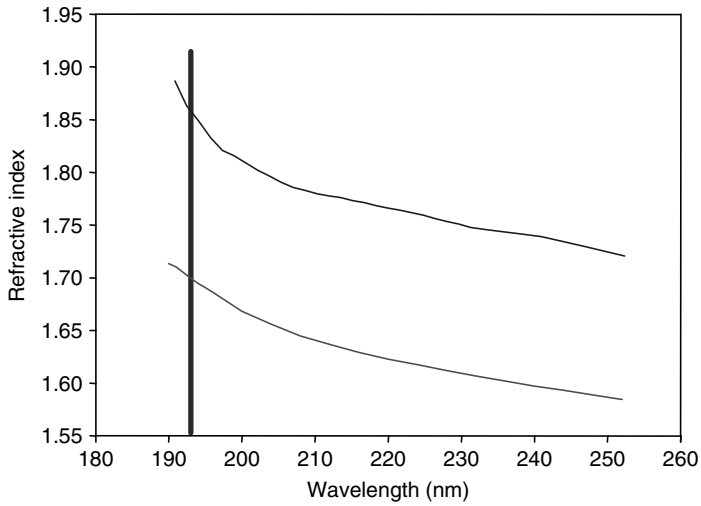


FIGURE 19.57 UV spectrograph of polymers from Figure 19.55 and Figure 19.56.

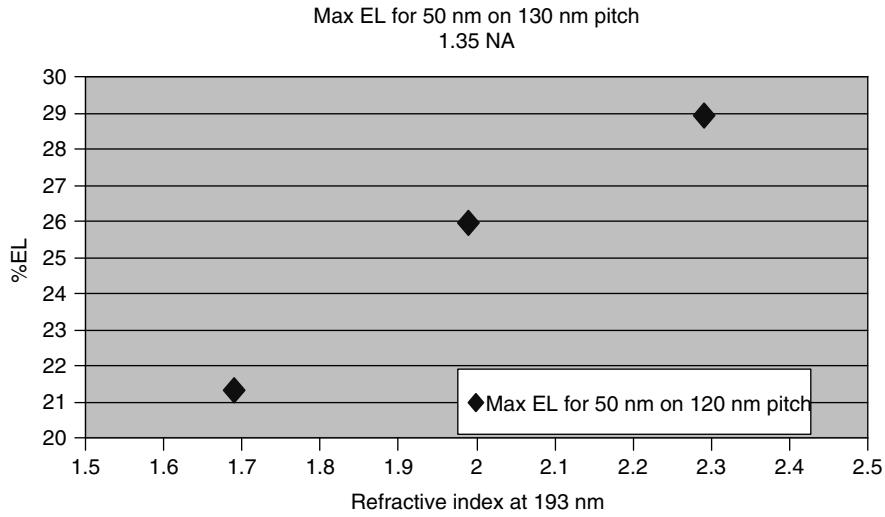


FIGURE 19.58 Exposure latitude vs. photoresist refractive index.

of isolated features,[146] requirements are that the resist will be used at a film thickness between 40 and 80 nm, will exhibit a LER of not greater than 1 nm per edge (3σ) and will support overall control of CDs to 1 nm (3σ).[112,145] These tolerances are smaller than the dimensions of the polymer molecules that constitute today's resists,[146] and given a typical carbon–carbon bond length of 0.13–0.15 nm,[147] it is clear that this specification is a call for atomic-scale control. To find practical use, a resist material must satisfy an extensive, comprehensive list of functional properties. Any viable resist must simultaneously achieve the target resolution, adequate sensitivity and acceptable imaging precision. These attributes ultimately

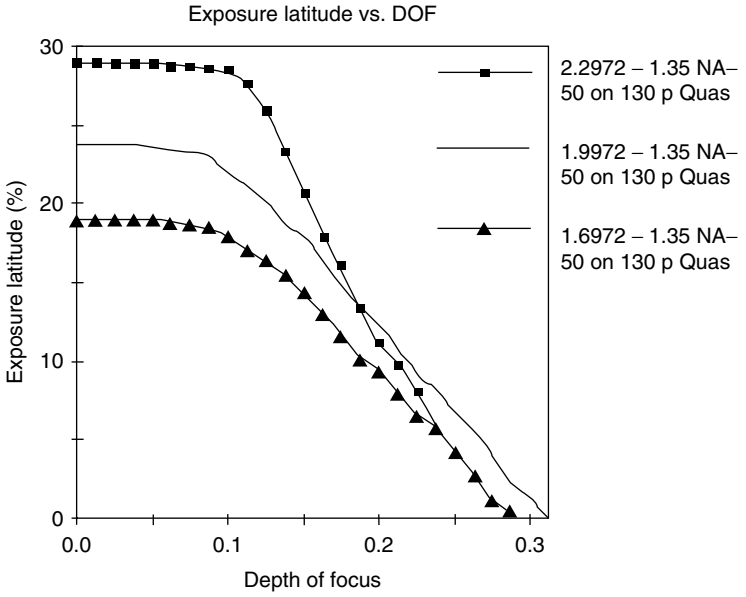


FIGURE 19.59 Process capability vs. refractive index.

are dictated by economics: the need to produce a product that the market wants at acceptable cost. Resist resolution determines the number of devices per circuit, device speed, and the number of devices per wafer; resist sensitivity governs wafer throughput per tool; and imaging precision affects device performance, and yield. Advanced research, largely carried out at academic laboratories active in nanoscience and nanofabrication, has sought to identify and extend the limits of nanoscale lithography. Among more conventional organic resist materials, the consensus is that PMMA is capable of imaging line-space arrays (formed by electron beam lithography) at the 15–20 nm scale (30–40 nm pitch)[147–153] without excessive LER,[124] currently the record for a polymer-based resist. Other nonpolymeric organic resist materials have been reported to exhibit similar resolution and low LER.[154,155] While these studies provide proof that resolution and LER consistent with the 2016-roadmap requirements is, in principle, attainable by currently known means, the radiation sensitivity of the materials used for these demonstrations is inadequate by orders of magnitude. The anticipated low brightness of NGL radiation sources is such that resists with very high radiation sensitivity will be required. Resist resolution criterion for 20-nm scale lithography development of EUV and EBP prototype tools.[144,156] The expectation that CA resists will be used with NGL is signaled by an ITRS specification of allowable change in image size with PEB temperature.[157] A key issue, then, and still unproven, is whether CA resists can simultaneously satisfy resolution and image precision specifications, while maintaining adequate radiation sensitivity.

References

1. Kunz, R. R., S. C. Palmateer, A. R. Forte, R. D. Allen, G. M. Wallraff, R. A. DiPietro, and D. Hofer. *Proc. SPIE* 2724 (1996): 365–76.
2. Garza, C. M., and W. L. Krisa. "Tools to Extend the Resolution of Optical Lithography." In *10th International Conference on Photopolymers*. Ellenville, NY: Society of Plastic Engineers, Oct. 1994, Mid-Hudson Section, Nov. 1994.
3. Thompson, L. F., C. G. Willson, and M. J. Bowden, eds. *Introduction to Microlithography, ACS Symposium Series 219*, 90–1. Washington, DC: American Chemical Society, 1983.
4. Thompson, L. F., C. G. Willson, and M. J. Bowden, eds. *Introduction to Microlithography, ACS Symposium Series 219*, 111–6. Washington, DC: American Chemical Society, 1983.
5. Meyerhofer, D. *IEEE Trans. Electron Devices* ED-27 (1980): 921.
6. Hinsberg, W. D., C. G. Willson, and K. K. Kanazawa. *Proc. SPIE* 539 (1985): 6.
7. Ouano, A. C. *Polym. Eng. Sci.* 18 (1978): 306.
8. Blum, L., M. E. Perkins, and A. W. McCullough. *Proc. SPIE* 771 (1987): 148.
9. Arcus, R. A. *Proc. SPIE* 631 (1986): 124.
10. Pampalone, T. R. *Solid State Technol.* 27, no. 6 (1984): 115.
11. Stillwagon, L. E. *Solid State Technol.* 28, no. 5 (1985): 113.
12. Hanabata, M., Y. Uetani, and A. Furuta. *Proc. SPIE* 920 (1988): 349.
13. Grant, B. D., N. J. Clecak, R. J. Twieg, and C. G. Willson. *IEEE Trans. Electron Devices* ED-28 (1981): 1300.
14. Templeton, M. K., C. R. Szmanda, and A. Zampini. *Proc. SPIE* 771 (1987): 136.
15. Pawloski, A. R., J. A. Torres, P. F. Nealey, and J. J. de Pablo. *J. Vac. Sci. Technol. B,* 17, no. 6 (1999).
16. Koshiba, M., M. Murata, M. Matsui, and Y. Harita. *Proc. SPIE* 920 (1988): 364.
17. Honda, K., B. T. Beauchemin Jr., E. A. Fitzgerald, A. T. Jeffries III., S. P. Tadros, A. J. Blakeney, R. J. Hurditch, S. Tan, and S. Sakaguchi. *Proc. SPIE* 1466 (1991): 141.
18. Kajita, T., T. Ota, H. Nemoto, Y. Yumoto, and T. Miura. *Proc. SPIE* 1466 (1991): 161.
19. Borzo, M., J. J. Rafalko, M. Joe, R. R. Dammel, M. D. Rahman, and M. A. Ziliox. *Proc. SPIE* 2438 (1995): 294.
20. Beauchemin, B. T., K. Honda, and R. J. Hurditch. *Proc. Electrochem. Soc., Patterning Sci. Technol.* 90-1 (1989): 15.
21. Uenishi, K., Y. Kawabe, and T. Kokubo. *Proc. SPIE* 1466 (1991): 102.
22. Trefonas, P. III., and B. K. Daniels. *Proc. SPIE* 771 (1987): 194.

23. Murata, M., M. Koshiba, and Y. Harita. *Proc. SPIE* 1086 (1989): 48.
24. With permission from Hinsberg W. D., et al. personal communication.
25. Reiser, A. *Photoreactive Polymers*, 179. New York: Wiley-Interscience, 1989.
26. Huang, J. P., T. K. Kwei, and A. Reiser. *Proc. SPIE* 1086 (1989): 74.
27. Ueberreiter, K., and F. Asmussen. *J. Polym. Sci.* 57 (1962): 187.
28. Asmussen, F., and K. Ueberreiter. *J. Polym. Sci.* 57 (1962): 199.
29. Moreau, W. M. *Semiconductor Lithography: Principles, Practices, and Materials*. 473. New York: Plenum Press, 1988.
30. Garza, C. M., C. R. Szmanda, and R. L. Fisher Jr. *Proc. SPIE* 920 (1988): 41.
31. Yeh, T.-F., H.-Y. Shih, and A. Reiser. *Macromolecules* 25 (1992): 5345.
32. Yeh, T.-F., A. Reiser, R. R. Dammel, G. Pawlowski, and H. Roeschert. *Macromolecules* 26 (1993): 3862.
33. Shih, H.-Y., T.-F. Yeh, A. Reiser, R. R. Dammel, H. J. Merrem, and G. Pawlowski. *Macromolecules* 27 (1994): 3330.
34. Shih, Y., and A. Reiser. *Macromolecules* 29 (1996): 2082.
35. Shih, H.-Y., and A. Reiser. *Macromolecules* 30 (1997): 3855.
36. Kim, M. S., and A. Reiser. *Macromolecules* 30 (1997): 3860.
37. Shih, H.-Y., and A. Reiser. *Macromolecules* 30 (1997): 4353.
38. Kim, M. S., and A. Reiser. *Macromolecules* 30 (1997): 4652.
39. Shih, H.-Y., H. Zhuang, A. Reiser, I. Teraoka, J. Goodman, and P. M. Gallagher-Wetmore. *Macromolecules* 31 (1998): 1208.
40. Stauffer, D., and A. Aharony. *Introduction to Percolation Theory*. London: Taylor and Francis, 1992.
41. Tsiartas, P. C., L. W. Flanagin, C. L. Henderson, W. D. Hinsberg, I. C. Sanchez, R. T. Bonnecaze, and C. G. Willson. *Macromolecules* 30 (1997): 4656.
42. Flanagin, L. W., C. L. McAdams, W. D. Hinsberg, I. C. Sanchez, and C. G. Willson. *Macromolecules* 32 (1999): 5337.
43. Flanagin, L. W., V. K. Singh, and C. G. Willson. *J. Polym. Sci., Part B: Polym. Phys.* 37 (1999): 2103.
44. Flanagin, L. W., V. K. Singh, and C. G. Willson. *J. Vac. Sci. Technol. B* 17 (1999): 1371.
45. Ito, H. *J. Photopolym. Sci. Technol.* 11 (1998): 379.
46. Kawamura, Y., T. Toyoda, and S. Namba. *J. Appl. Phys.* 53 (1982): 6489.
47. Allen, R. D., G. M. Wallraff, W. D. Hinsberg, and L. L. Simpson. *J. Vac. Sci. Technol.* B9 (1991): 3357.
48. Kaimoto, Y., K. Nozoaki, S. Takechi, and N. Abe. *Proc. SPIE* 1672 (1992): 66.
49. Gipstein, E., A. C. Ouano, and T. Thompkins. *J. Electrochem. Soc.* 129 (1982): 201.
50. Hanrahan, M. J., and K. S. Hollis. *Proc. SPIE* 771 (1987): 128.
51. Pawalshi, T., T. Sauer, R. Dammel, D. J. Gordon, W. Hinsberg, W. McKean, C. Lindler, H. Merrem, R. Vicari, and C. G. Willson. *Proc. SPIE* 1262 (1990): 391.
52. Reichmanis, E., C. W. Wilkins, and E. A. Chandross. *J. Vac. Sci. Technol.* 19 (1981): 1338.
53. Swartzkopf, G., K. B. Gabriel, and J. B. Covington. *Proc. SPIE* 1262 (1987): 456.
54. Houlihan, et al. *Proc. SPIE* 2195 (1991): 231.
55. Iwayanagi, T., T. Kohashi, S. Nonogaki, T. Matsusawa, K. Douta, and H. Yanazawa. *IEEE Trans. Electron Devices* ED-25 (1981): 1306.
56. Lin, B. J. *J. Vac. Sci. Technol.* 12 (1975): 1317.
57. Nate, K., and T. Kobayashi. *J. Electrochem. Soc.* 128 (1981): 1394.
58. Ito, H., and C. G. Willson. In *Polymers in Electronics*, edited by T. Davidson, 11. Washington, DC: American Chemical Society, 1984.
59. Crivello, J. V. In *UV Curing: Science and Technology*, edited by S. P. Pappas, Norwalk, CT: Technology Marketing Corp., 1978.
60. Ito, H., and C. G. Willson. *Polym. Eng. Sci.* 23 (1983): 1012.
61. Feeley, W. E., J. C. Imhof, C. M. Stein, T. A. Fischer, and M. W. Legenza. *Polym. Eng. Sci.* 26 (1986): 1101.
62. Thackeray, J. W., G. W. Orsula, E. K. Pavelcheck, and D. Canistro. *Proc. SPIE* 1086 (1989): 34.
63. Ito, H., J. M. Frechet, and C. G. Willson. U.S. Patent.

64. Wallraf, G. M., W. D. Hinsberg, F. Houle, J. Opitz, D. Hopper, and J. M. Hutchinson. *Proc. SPIE* 2438 (1995): 182.
65. Crivello, J. V. In *Polymers in Electronics*, edited by T. Davidson, 3. Washington, DC: American Chemical Society, 1984.
66. Crivello, J., and J. Lam. *J. Polym. Sci.* 16 (1978): 2441.
67. McKean, D. R., U. Shaedeli, and S. A. Macdonald. In *Polymers in Microlithography*, edited by E. Reichmanis, 27. Washington, DC: American Chemical Society, 1989.
68. Crivello, J. In *Polymers in Electronics*, edited by T. Davidson, ACS Symposium Series, 242. Washington, DC: American Chemical Society, 1984.
69. Brunsvold, W., R. Kwong, W. Montgomery, W. Moreau, H. Sachdev, and K. Welsh. *Proc. SPIE* 1262 (1990): 162.
70. Hacker, N., and K. Welsh. *Proc. SPIE* 1466 (1991): 384.
71. Iwamoto, T., S. Nagahara, and S. Tagawa. *J. Photopolym. Sci. Technol.* 11 (1998): 455.
72. Renner, C. U.S. Patent 4,371,605, 1983.
73. Kasai, P. H. *JACS* 114 (1992): 2875.
74. Szmanda, C. R., R. Kanangh, J. Buhland, J. Cameron, P. Trefonas, and R. Blacksmith. *Proc. SPIE* 3678 (1999): 857.
75. Poot, A., G. Delzene, R. Pollet, and U. Laridon. *J. Photogr. Sci.* 19 (1971): 88.
76. Pawlowski, G., R. Dammel, C. R. Lindley, H. Merrem, H. Roschert, and J. Lingnau. *Proc. SPIE* 1262 (1990): 16.
77. Houlihan, F. M., A. Shugard, R. Gooden, and E. Reichmanis. *Macromolecules* 21 (1988): 2001.
78. Houlihan, F. M., E. Chin, O. Nalamasu, J. M. Kometani, T. X. Neenan, and A. Pangborn. *Proc. SPIE* 2195 (1994): 137.
79. Dill, F. H., W. P. Hornberger, P. S. Hauge, and J. M. Shaw. *IEEE Trans. Electron Devices* ED22 (1975): 445.
80. Sturtevant, J. IBM Technical Report TR-19.0938, 1991.
81. McKean, D., U. Schaedeli, and S. J. Macdonald. *J. Polym. Sci. Chem. Ed.* 27 (1989): 3927.
82. Sturtevant, J. S., W. Conley, and S. E. Webber. *Proc. SPIE* 2724 (1996): 273.
83. Tarascon, R., E. Reichmanis, F. M. Houlihan, H. Shugard, and L. F. Thompson. *Polym. Eng. Sci.* 29 (1989): 850.
84. Turner, S. R., K. D. Ahn, and C. G. Willson. In *Polymers for High Technology*, edited by M. J. Bowden, and S. R. Turner, 200. Washington, DC: American Chemical Society, 1984.
85. Przybilla, R. J., R. Dammel, H. Roshert, W. Spiess, and G. Pawlosky. *J. Photopolym. Sci. Technol.* 4 (1991): 421.
86. Ito, H., G. Breyta, R. Sooriyakumaran, and D. Hofer. *J. Photopolym. Sci. Technol.* 8 (1995): 505.
87. Nozaki, K., K. Watanabe, E. Yano, A. Kotachi, S. Takechi, and I. Hanyu. *J. Photopolym. Sci. Technol.* 9 (1996): 509.
88. Szmanda, C., R. Kavanagh, P. Trefonas, and R. Blacksmith. *Proc. SPIE* (1998).
89. Smith, G. H., S. Paul, and J. A. Bonham. U.S. Patent 3,779,778, 1973.
90. Hayashi, N., S. Hesp, T. Ueno, M. Toriumi, T. Iwayanagi, and S. Nonogaki. *Polym. Mater. Eng.* 61 (1989): 417.
91. Hayashi, N., L. Schlegel, T. Ueno, H. Shiraishi, and T. Iwayanagi. *Proc. SPIE* 1466 (1991): 377.
92. Mertesdorf, C., N. Munzel, H. Holzarth, P. Falcigno, H. Schact, O. Rohde, S. Schulz, et al. *Proc. SPIE* 2438 (1995): 84.
93. Dossel, K. F. EP-Appl 0 312 751, 1988.
94. Bantu, et al. *Proc. SPIE.*
95. Lee, K. Y., and W. S. Huang. *J. Vac. Sci. Technol. B* 11 (1993): 2807.
96. Conley, W., et al. "SEMATECH DUV Workshop." Austin, TX, Nov. 1992.
97. Cameron, et al. *Proc. SPIE* 4345 (1993): 106.
98. Sundararajan, et al. *Proc. SPIE* 3678 (1992): 78.
99. Michaelson, et al. *Proc. SPIE* 5753 (1992): 368.
100. Lewis, C., and C. G. Willson. *Proc. SPIE* (1999): 3678.
101. Mack, C. A. *J. Electrochem. Soc.* 134 (1987): 148.

102. Ito, H., D. F. Alexander, and G. Breyta. *J. Photopolym. Sci. Technol.* 10 (1997): 397.
103. Ito, H., and E. Flores. *J. Electrochem. Soc.* 135 (1988): 2322.
104. Itani, T., H. Yoshino, S. Hashimoto, M. Yamana, N. Samoto, and K. Kasama. *J. Photopolym. Sci. Technol.* 10 (1997): 409.
105. Itani, T., H. Iwasaki, H. Yoshino, M. Fujimoto, and K. Kasama. *Proc. SPIE* 2438 (1995): 191.
106. Iwasa, S., K. Maeda, K. Nakano, and E. Hasegawa. *Proc. SPIE* 3049 (1997): 126.
107. Yamachika, M., K. Patterson, J. D. Byers, and C. G. Willson. *J. Photopolym. Sci. Technol.* (1999): 12.
108. Allen, R. D., G. M. Wallraff, W. D. Hinsberg, L. L. Simpson, and R. R. Kunz. In *Polymers for Microlelectronics*, edited by L. F. Thompson, C. G. Willson, and S. Tagawa, 165. Washington, DC: American Chemical Society, 1994.
109. Conley, W., B. Brunsvold, F. Buehrer, R. Dellaguardia, D. Dobuzinsky, T. Farrel, H. Ho, et al. *Proc. SPIE* 3049 (1997): 282.
110. Aoai, T., T. Yamanaka, and M. Yagihara. *J. Photopolym. Sci. Technol.* 10 (1997): 387.
111. Barclay, G. G., C. J. Hawker, H. Ito, A. Orellana, P. R. L. Malenfant, and R. F. Sinta. *Proc. SPIE* 2724 (1996): 249.
112. "International Technical Roadmap for Semiconductors." Austin, TX: SEMATECH, Inc., 2001.
113. "Special Issue on Limits of Semiconductor Technology." *Proc. IEEE* 89, no. 3 (2001).
114. Ito, T., and S. Okazaki. *Nature* 406 (2000): 1027.
115. Harriott, L. ref. 2, 366–374.
116. Allen, et al. U.S. Patent 5,071730.
117. Allen, R. D., G. M. Waliraff, R. A. DiPietro, and D. C. Hofer. "193 nm Single Layer Positive Resists Building Etch Resistance into a High Resolution Imaging System." *Proc. SPIE* 474 (1995): 2438.
118. Khojasteh, M., K. Chen, R. Kwong, M. Lawson, P. Varanasi, K. Patel, and E. Kobayashi. "High-Performance 193-nm Photoresist Materials Based on a New Class of Polymers Containing Spaced Ester Functionalities." *Proc. SPIE* 5039 (2001): 187.
119. Hada, et al. *Proc. SPIE* 5039 (2001): 752.
120. Jonathan, C., et al. *Proc. SPIE* 5039 (2003): 376.
121. Lee, J. Y., et al. *Proc. SPIE* 5376 (2004): 426.
122. Stewart, D. M., et al. *Proc. SPIE* 5039 (2003): 415.
123. Bryan, J. R., et al. *Proc. SPIE* 5039 (2003): 376.
124. Yoshizawa, M., et al. *Proc. SPIE* 3997 (2000): 301.
125. Yoshizawa, M., et al. *J. Vac. Sci. Technol. B* 19, no. 6 (2001): 2488.
126. Allen, R., et al. *Proc. SPIE* 5753 (1995): 256.
127. Hung, R., et al. *Proc. SPIE* 4345 (1993): 385.
128. Trinque, et al. *Proc. SPIE* 4690 (2003): 58.
129. Conley, W., P. Zimmerman, D. Miller, and G. S. Lee. *Proc. SPIE* 5039: 207.
130. Varanasi, P. R., R. W. Kwong, M. Khojasteh, K. Patel, K.-J. Chen, W. Li, and M. C. Lawson, et al. *Proc. SPIE* 5753 (2004): 131.
131. Lin, B. J. "The k3 Coefficient in Nonparaxial λ/NA Scaling Equations for Resolution, Depth of Focus, and Immersion Lithography." *Microlithography, Microfabrication, Microsystems* 3, no. 7, (2002).
132. Abbe, E. "Beiträge zur Theorie des Midroskops und der mikroskopischen Wahrnehmung." *Archiv für Mikroskopische Anatomie* (1873).
133. Conley, W., et al. "SEMATECH Immersion Lithography Workshop-IBM Almaden Research Centre." San Jose, CA, Jul. 2003.
134. Conley, W., et al. "1st International Symposium on Immersion Lithography." Vancouver, BC, Aug. 2004.
135. Pollentier, I., et al. *Proc. SPIE* 5754 (2001): 129.
136. Dammel, R., et al. *Proc. SPIE* 5753 (2001): 95.
137. Webb, J. E. "Extending the Newtonian Design Form for Ultra-High Numerical Aperture and Immersion Lithography." *Proc. SPIE* 5377 (2002): 69.

138. Mulkens, J., D. Flagello, B. Streefkerk, and P. Graupne. "Benefits and Limitations of Immersion Lithography." *JM3* January, (2003).
139. Kusumoto, S., et al. *Proc. SPIE* 5753: 10; Peng, S., et al. *Proc. SPIE* 5754 (2004): 427.
140. Dammel, R. *J. Photopolym.* June, (2005).
141. Khojastch, M., K. Chen, R. Kwong, M. Lawson, P. Varanasi, and P. Patel. *Proc. SPIE* 5039 (1873): 187.
142. Whittaker, A. K., I. Blakey, H. Liu, D. J. T. Hill, G. A. George, W. Conley, and P. Zimmerman, "High-RI Resist Polymers for 193 nm Immersion Lithography." *Proc. SPIE* 5753 (2003): 827.
143. Conley, W., and J. Bendik. "Is ArF the Final Wavelength?" *Proc. SPIE* 5376 (2004): 16.
144. Bjorlkholm, S. *Intel Technol. J.* 3rd Quarter 1998, available at http://www.intel.com/technology/itj/q31998.htm
145. Levinson, H. *IEEE Circuits Dev. Mag.* 18 (2002): 50.
146. Broers, A., A. Hoole, and J. Ryan. *Microelectron. Eng.* 32 (1996): 131.
147. Castellan, G. *Physical Chemistry.* 2nd ed., 578. Reading, MA: Addison-Wesley, 1971.
148. Billmeyer, F. *Textbook of Polymer Science.* 2nd ed., 154–7. New York: Wiley-Interscience, 1984.
149. Issacson, M., and A. Murray. *J. Vac. Sci.Technol.* 19 (1981): 1117.
150. Broers, A. *J. Electrochem. Soc.* 128 (1981): 166.
151. Vieu, C., F. Carcenac, A. Pepin, Y. Chen, M. Mejias, A. Lebib, L. Manin-Ferlazzo, L. Courad, and H. Launis. *Appl. Surf. Sci.* 164 (2000): 111.
152. Yasin, S., D. Hasko, and H. Ahmed. *Microelectron. Eng.* 61-62 (2002): 745.
153. Chen, W., and H. Ahmed. *Appl. Phys. Lett.* 62 (1993): 1499.
154. Fujita, J., Y. Onishi, Y. Ochiai, and S. Matsui. *Appl. Phys. Lett.* 68 (1996): 1297.
155. Robinson, A., R. Palmer, T. Tada, T. Kanayama, M. Allen, J. Preece, and K. Harris. *J. Phys. Dev. Appl. Phys.* 32 (1999): L75.
156. Dhaliwal, R., W. Enichen, S. Golladay, M. Gordon, R. Kendall, J. Lieberman, H. Pfeiffer, et al. *IBM J. Res. Dev.* 45 (2001): 615.
157. Chumanov, G., D. D. Evanoff Jr., I. Luzinov, V. Klep, B. Zdyrko, W. Conley, and P. Zimmerman. "Nanocomposite Liquids for 193 nm Immersion Lithography: A Progress Report." *Proc. SPIE* 5753 (2003): 847.

20

Photomask Fabrication

20.1 Introduction **20**-1

20.2 Photomask: Structure and Fabrication **20**-2
Evolution of the Photomask

20.3 Writing Patterns on Masks **20**-19
Mask Writers and Data Preparation

20.4 Materials and Processing **20**-39
Glass • Chrome • Molybdenum Silicide • Other Substrate Related Topics • Photoresist • Processing • Pellicles

20.5 Photomask Qualification **20**-47
Metrology • Inspection and Repair • Phase, Transmission, and Image Evaluation

20.6 Manufacturability and COO **20**-67
Photomask Yield • Cycle Time to Mask User • On-Time Delivery Performance • Cost • Cost of Ownership • Device Specific Reticle Issues

References **20**-71

Syed A. Rizvi
Nanotechnology Education and Consulting Services

Sylvia Pas
Texas Instruments, Inc.

20.1 Introduction

The use of photomasks started with the birth of semiconductor technology during the second half of the last century when discrete devices were first developed. As those devices evolved into ICs and later into microprocessors the complexities of photomasks grew. A major contribution to the advancements in semiconductor technology comes from the field of lithography where photomasks have played a key role. In fact, photomasks are to be regarded as one of the constituents of the field of lithography, like photoresist, light sources, and imaging-optics are.

The focus of lithography in semiconductor processing has been on the printing of features on wafers. When the features became very small where they could be measured in microns rather than in mils (one thousandth of an inch) the term "lithography" was replaced by a more appropriate term "microlithography," and now for obvious reasons industry is beginning to use the term "nanolithography." However, the term "photolithography" is also commonly used in the industry.

Basically, a photomask is a glass plate with some kind of pattern to be copied onto a wafer previously coated with a photosensitive film. The imaging of the pattern on the wafer is carried out by exposing the photosensitive film to light transmitted through the photomask and then developing the resist.

It is because of the use of light in the process of imaging that the terms like photolithography and photomask came to be used. In recent years other means of exposures like electron-beam, ion-beam, and x-rays have also been introduced where these technologies are referred as next generation lithography (NGL). On occasions, depending on the context of the material, NGL is referred as post-optical lithography while the conventional photolithography is called optical-lithography. In any case, the term

"photomask" continues to be used for both optical and non-optical lithography. Moreover, the term photomask is quite often referred as simply "mask" for brevity.

Among many other factors the wavelength of light, used for the imaging process, also has an impact on the miniaturization of feature that can be printed on a wafer. The smaller the wavelength, the smaller the feature can be printed. In order to meet the demands for continually shrinking features the lights used for the lithographic systems have gone through several generations of wavelength reductions, namely, near-ultra violet (436 nm), ultraviolet (UV) (365 nm), and now deep-ultraviolet (DUV) (248, 193, and 157 nm). Significant progress has also been made in extending the technology down to extreme ultraviolet (EUV) (13.6 nm) but because of its x-ray like properties it is placed under the class of NGL. With the introduction of NGL came along NGL masks that differ significantly from their optical counterpart as regards to their structure but basically they all have one thing in common that they all have some kind of patterns on them that are to be imaged on wafers.

This chapter highlights the current advancements in the area of optical photomasks. For a comprehensive knowledge on optical as well NGL photomask the reader is encouraged to look into the Handbook of Photomask Manufacturing Technology, also Published by Taylor & Francis, in 2005.

20.2 Photomask: Structure and Fabrication

The structure and hence the fabrication of a photomask can be classified under matured-mask technologies and advanced-mask technologies, although the line between the two has been shifting as advanced technologies are becoming more matured with the passing of time. This classification between matured and advance types is based on the actual structure of a photomask rather than on the advancement in the processes and equipments used in their fabrication.

The basic structure of a photomask, whether matured or advanced, has remained very much unchanged since its inception. Except for the early years of semiconductor technology when the masks used to be emulsion-on-glass masks, all masks since then have been chrome-on-glass (COG) masks where the presence or absence of chrome features on glass defines opaque and transparent regions representing a circuit-layout to be printed on wafers. With continued shrinkage of features the demands on the tolerance of feature-width and feature-placement have been growing. Among lithographers these features are referred as critical dimension (CD), and the related tolerance are called CD and image placement (IP) control.

20.2.1 Evolution of the Photomask

In the following sections, we review the evolution in usage that brought about the changes in the mask itself, and led to matured and advanced technology addressed later in the chapter.

20.2.1.1 Use of a Generic Mask on Wafer

As mentioned earlier, a photomask is an integral part of the overall lithography process. Figure 20.1 [1] shows the role of the photomask in a lithographic process flow. The starting material in the process is a silicon wafer on which a film (oxide or nitride) is previously grown. The wafer is then coated with a photosensitive material known as photoresist or simply resist. The next step is to expose the photoresist to some kind of light that is transmitted through a mask such that the image of a pattern on the mask is formed on the resist surface. The photoresist is then developed which then reproduces the mask pattern on the resist film. This pattern is then transferred on to the oxide/nitride film on wafer by etching off parts of the film not protected by the resist. Finally all resist is removed leaving behind an etched pattern on wafer that forms a replica of the pattern on the original mask.

20.2.1.1.1 Contact Printing and Proximity Printing

In the early days of photolithography the masks used to be put into intimate contact with resist coated wafers. This technique was referred as 1× imaging or 1:1 imaging, since, unlike today's process, the printed images on the wafers used to be of the same size as on masks. The problem with this approach

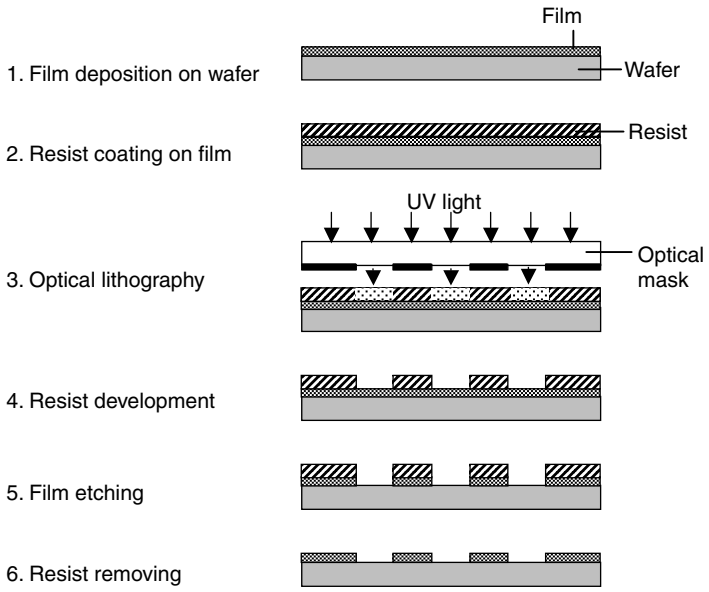


FIGURE 20.1 Process flow for lithography. (From Yoshioka, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 135–56. Boca Raton, FL, 2005.)

was that after only few exposures, some of the resists on wafers would get adhered to the masks, which then needed frequent cleaning. This process also used to cause mask damage because of its repetitious contacts with resist coated wafers. A solution to the problem was found by leaving a small gap between the mask and the resist film and this technique was called proximity printing. Although proximity printing did solve the resist adherence problem, the process also resulted in the degradation of image quality. Examples of contact-printing and proximity-printing are shown in Figure 20.2.

20.2.1.1.2 Projection Printing (1×)

Proximity printing although increased the mask life but the technique was not able to deliver the resolution and feature definition that were becoming more demanding as the features were getting smaller with each generation of devices.

In order to obtain both, high resolution and longer mask life, a new technique known as projection printing was then introduced. Earlier techniques for 1:1 projection printing employed refractive optics consisting of assembly of lenses, but lenses with large diameters capable of imaging the entire array of patterns covering the whole mask in one single shot (exposure) could not be built without inherent aberrations associated with lenses. Moreover, industry was also gearing up for larger wafers that required

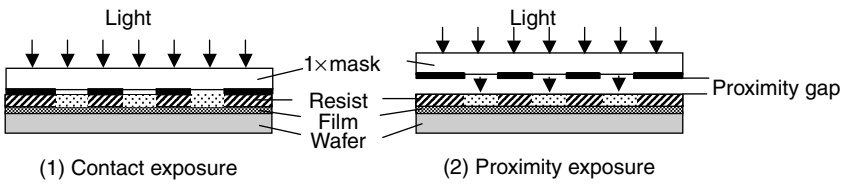


FIGURE 20.2 Contact print and proximity print at 1×. (From Yoshioka, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 135–56. Boca Raton, FL, 2005.)

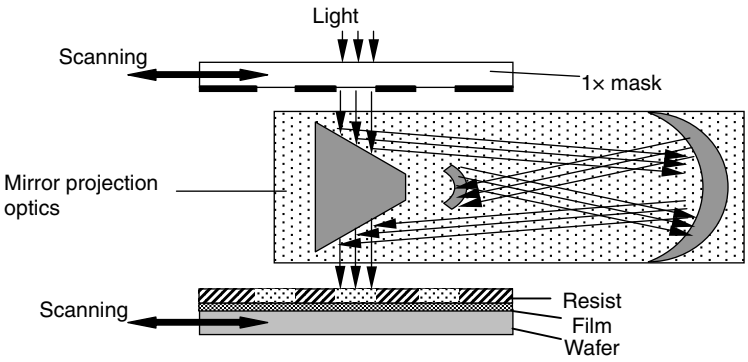


FIGURE 20.3 1×Projection printing using reflective optics. (From Yoshioka, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 135–56. Boca Raton, FL, 2005.)

lenses with even larger diameters for 1:1 printing. In the end the 1:1 printing with refractive optics could not keep up with new challenges.

The next approach was then to employ mirrors along with a scanning mechanism. This technique worked for almost a decade. The absence of refractive optics overcame image degradation and lens aberration problems. Figure 20.3 shows a schematic of the optics of such a system. Because of the 1× masks the specs on the feature tolerance on masks remained same as they were for wafers. The lithography tolerance on mask features depended on process capability of mask fabrication that put significant constraints on the limitation of 1:1 projection optics.

20.2.1.1.3 Reduction-Projection Printing (10×, 5×, 4×, etc.)

The above problem was then addressed by introducing reduction-projection optics as shown in Figure 20.4 In this method, masks with larger features (4×, 5×, or even 10×) were employed for printing on wafers. Specs on larger features were more forgiving since small errors on mask when printed

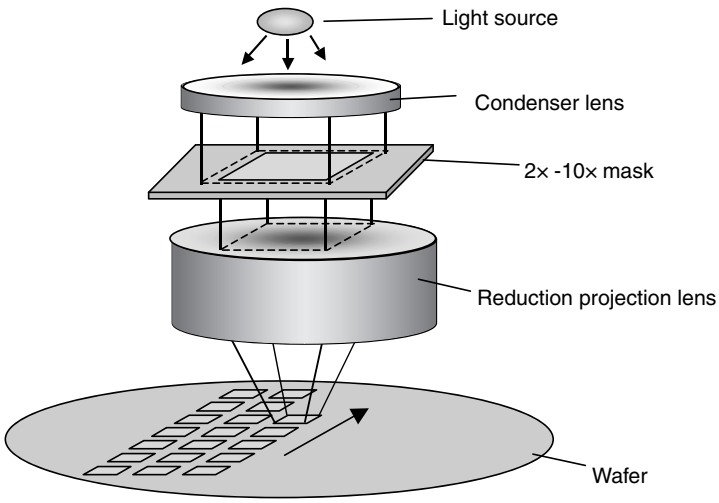


FIGURE 20.4 Reduction projection printing using refractive optics. (From Yoshioka, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 135–56. Boca Raton, FL, 2005.)

on wafer would be reduced to the point of extinction. This reduction-projection printing worked well for quite some time.

The masks used for reduction-projection printing were different from their 1× predecessors in a significant way. Whereas, the earlier masks used to have a large array of 1× dies the later days reduction-projection masks carried only a few 4× or 5× dies per mask. As a matter of fact, when the idea of reduction-projection was first introduced, masks used to have only one 10× die. Later on 4× and 5× dies were introduced to accommodate for more dies per mask in order to increase the throughput of the system.

20.2.1.1.3.1 Photomasks Used in Reduction Projection Printing

A schematic structure of a 6-in optical mask, used in reduction optics is shown in Figure 20.5. The photomask is made of glass (quartz) substrate with chrome (Cr) pattern on it. In order to protect the pattern from contamination and foreign particles the mask is mounted with a pellicle, also shown in the figure. The earlier masks with an array of dies, in form of a matrix with rows and columns, were made by what was known as step and repeat operation where another mask known as reticle, with a 10× magnified image of one single die, was printed on mask using 10× reduction optics; the operation was then repeated to form an array of dies on the entire mask.

The reticle on the other hand, was made by exposing a photo-sensitive plate under a tool known as pattern generator (PG). This class of tools was equipped with a computer controlled stage that allowed for the movement of the reticle plate in the *X* or *Y* direction. An aperture in the optical path allowed for a variable shaped slit from a wide range of height and width values. The optics and aperture were mounted on a turret that could be rotated in 0.1° increments and allowing for angled features as well. These machines created a 10× magnified replica of the die layout on the photosensitive plate that was called reticle. During the later years, the PGs were replaced by electron beam writers.

In the mid 1980s the printing of the entire mask on wafer in one single shot began to be replaced by stepping the reticle directly on to wafers, thus bypassing the need for printing an array on a mask and then printing of the mask on a wafer. This was also the time when the distinction between mask and reticle began to be blurred because these reticles had replaced the function of masks and for that reason many called these reticles as masks. Even today both terms "reticles" and "masks" are used interchangeably. Later on, in order to increase the throughput of the system 10× reticles were replaced by 5× and 4× reticles that accommodated more dies per reticle.

Masks with 10×, or even with 5× and 4× magnification gave an edge to the mask makers by loosening the tolerance in mask specification. This period of relaxed tolerance however was short lived not only because demands on the feature size continued to grow but also because new extremely small features were added to the existing patterns; even when those new features were not part of circuit and were not designed to be printed on wafers. This kind of innovation in mask design, among many others, opened a new chapter the way masks were designed which leads to topic of Advanced Mask Technology.

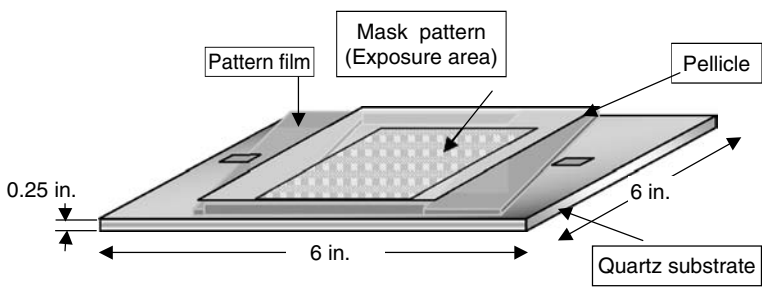


FIGURE 20.5 A photomask mounted with pellicle. (From Yoshioka, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 135–56. Boca Raton, FL, 2005.)

20.2.1.2 Advance Technologies for Making Optical Masks

20.2.1.2.1 Simple Theory of Imaging

Today's imaging systems require sophisticated optics, and it is therefore important to understand, at least in principle, the contributions from the various parameter that influence quality of images projected on wafer surface.

One of the frequently cited formulae relating to feature size to other imaging parameters is described as:

$$\text{Minimum feature} = k_1 \frac{\lambda}{\text{NA}}$$

The relationship states that the minimum feature defined as half of the smallest pitch to be printed is directly proportional to the wavelength λ of the illuminating light and inversely proportional to the numerical aperture NA of the lens. The expression (λ/NA) is referred as optical resolution and is a measure of the smallness of the feature that can be printed on wafer.

The value of the proportionality constant k_1 in the above equation depends on a number of factors that influence the resolution of the feature. In early days the value of k_1 used to be 0.61. The number was derived solely from the optical theory of image formation that involved the use of Bessel's Function encountered in equations dealing with circular aperture. In fact, the value of Bessel's function of the fist kind denoted by J_1 happens to be 1.22…, which is two times the value of $k_1 (= 0.61)$. It is thus obvious from the above equation that k_1 can be seen as measure of the smallness of feature that can be printed. In addition to the parameter like wavelengths and numerical apertures there are other factors like off-axis illumination (OAI), or the resist processing, etc., that can be optimized to print smaller and smaller features which can then make k_1 smaller even when the wavelength and numeral aperture remain unchanged. This area is also referred as low k_1 lithography.

The optical resolution defined by (λ/NA) has been improved by continued reduction in wave length from 436 nm down to 248 nm and now 193 nm. Although significant effort has been made to work with 157 nm, the material problem remains to be a show stopper at this wavelength. The parameter NA, which is directly proportional to the diameter of lens used to be 0.5 during the 1980s and now it's close to 0.85. Although increase in NA can lead to better resolution there is a down side to it; the depth of focus is inversely proportional to the square of NA, and therefore higher NA results in loss of depth of focus. Hence while designing a lithographic system these factors have to be taken into consideration. Table 20.1 enlists relationships among the various parameters at different values of k_1 [2].

For the basic optical concept used in lithography exposure a k_1 of 0.25 is generally regarded as the theoretical limit [3]. Beyond this point the image contrast and fidelity degrades very rapidly. During the last decades many strategies like Retical Enhancement Technologies (RET) or optical proximity corrections (OPC) have been developed and have been a subject of extensive research [4]. Retical Enhancement Technologies and OPC are regarded as the foundation of today's the Advanced Mask Technology.

Besides loss of resolution and poor fidelity when printing small features, there is another issue known as non-linearity that did not use to be a matter of concern in early days when the features did not use to be as small as they are today.

In the case of a perfect linear process any change in feature's dimension in mask would correspond to an identical change at wafer scale (after taking care of the demagnification factor of course). In a non-linear process on the other hand, a small change in mask dimension can be amplified into a significant change at the wafer. This phenomenon has an adverse impact on mask fabrication because any error in mask can result in a magnified error on wafer. The factor by which an error is magnified is known as mask error enhancement factor (MEEF) and defined by the following relationship [5].

TABLE 20.1 k_1 For Different Semiconductor Device Generations

Device Generation (nm halfpitch)	Exposure Wavelength (nm)	Numerical Aperture of Stepper Lens	k_1
500	365	0.5	0.68
250	248	0.5	0.5
200	248	0.6	0.48
180	248	0.63	0.46
140	248	0.75	0.42
110	193	0.75	0.43
90	193	0.85	0.40
70	157	0.75	0.33
50	157	0.85	0.27

Source: From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, ed. Rizvi, S., Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.

$$\text{MEEF} = \frac{\delta(\text{wafer linewidth})}{\delta(\text{mask linewidth})}$$

where δ means change in linewidth for mask or wafer.

It is apparent that in an ideal case MEEF would be 1.0. It is also obvious from above that a process with MEEF greater than 1 puts serious constraint on mask specification, on the other hand, when the MEEF is less than 1.0, the tolerance on mask specs would be more relaxed.

20.2.1.2.2 Two Areas of Advanced Mask Technology

The advanced mask technology can be classified under two groups as described in the following.

1. Under one group it is the physical structure of mask that is modified to compensate for the diffraction effect which is a major contributor to the loss in resolution and contrast. The masks from this group are called phase shift masks or simply PSMs.
2. Under the second group it is not the physical structure but rather the pattern on the conventional COG masks that are modified to compensate for the diffraction effect. In fact these masks can be regarded as an extension and refinement on the current GOG mask technology.

Later we will see that under the class of PSMs the masks can be further divided into sub-groups of PSMs. One sub-group requires etching into the substrate of a COG mask known as alternating PSM (Alt.PSM), and the other sub-group uses different materials to induce phase shift in light and known as halftone PSM (HT.PSM). We will see that mask under the second group can also be divided into sub-groups of masks with OPC, and masks with sub-resolution assist features (SRAF). These classifications are shown in the following. Since there is no agreed upon terminology that combines OPC and SRAF under one single name, both terms are being shown here in the following definition.

1. PSMs (Alt.PSM and HT.PSM).
2. Optical proximity corrections and SRAF.

In the following section we will address the PSMs.

20.2.1.2.3 Alt.PSM and HT.PSM

Under the class of PSMs we have Alt.PSM and HT.PSM. In the first part of this section we will address Alt.PSM.

20.2.1.2.3.1 Alternating Phase Shift Mask

Alternating PSMs are basically the conventional COG masks except that in these masks at certain specified locations the clear features of glass are etched to a predetermined depth. Thus, PSMs not only

change the intensity of light transmitted through the clear areas but also causes a phase change in the transmitted light.

Alt.PSM, also known as alternating aperture phase shift mask (AAPSM), improves the process window of a narrow dark feature by providing a 180° phase transition between two bright features defining this dark feature.

Figure 20.6 shows the basics of Alt.PSM. The left and right side pictures in this figure show the profiles of the electric field vector and of the intensity of light after it passes through the two openings separated by an opaque line. The two openings are such that the lights being transmitted through them are 180° out of phase with each other. This is shown by the solid lines in the two pictures. For the sake of comparison the dotted lines in the pictures show when there is no phase change. The impact of phase change on the resolution can clearly be seen in the picture on the right that shows the intensity profile corresponding to the solid and dotted lines. Since the intensity is represented by the square of the amplitude of the electric field the intensity under the opaque line is shown to be very close to zero. Such is not the case with dotted lines that represent no phase change taking place. The phase change thus can give an excellent resolution in imaging a pattern on wafer.

Besides the degree of resolution between the two cases there is also a phenomenon of printability as a function of size of the opaque feature. Figure 20.7 illustrate the two cases for "with" and "without" phase shift on its left and right pictures. In the cases of phase change, the intensity between the two opening remains zero regardless of how small the opaque line is, as is shown on the left picture. The right side picture where no phase change is involved, the minimum intensity increases as the opaque feature become smaller.

The printability relationship with and without phase shift is further explored in Figure 20.8. In case of phase shift mask the printed feature reaches a constant minimum value as the mask CD decreases as shown by a solid line. When phase shift is not involved the opaque feature size reduces to zero very rapidly as the mask feature becomes smaller.

20.2.1.2.3.2 Fabrication of Alt.PSM

Fabrication of most Alt.PSMs starts from COG mask followed by selective etching of quartz as prescribed by the design [6]. Figure 20.9 shows some basic steps involved in the fabrication of an Alt.PSM. As it can

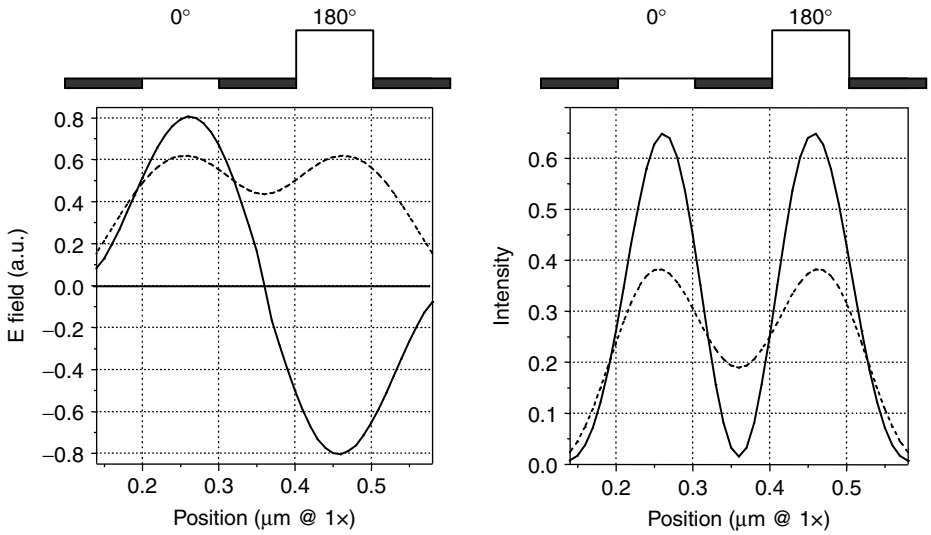


FIGURE 20.6 Aerial image simulation for two openings 100 nm each with 100 nm chrome line. 193 nm illumination, 0.75 NA and coherence factor 0.3. (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

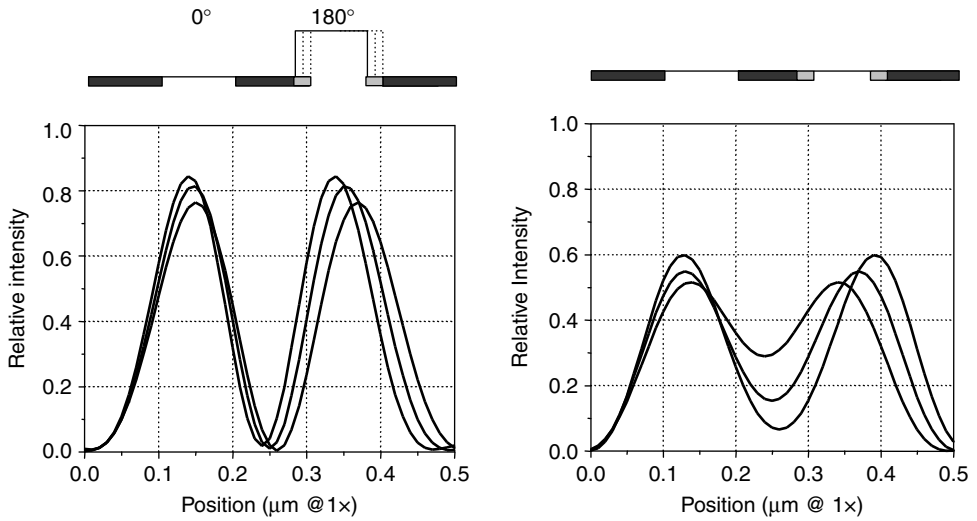


FIGURE 20.7 Aerial image simulation showing intensity distribution for alternating phase-shift mask (Alt PSM) (left) and conventional COG (right). 193 nm illumination, 0.75 NA and coherence factor 0.3. (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

be seen from the figure, that Alt.PSM will require at least two lithography steps. The first litho step is to etch through chrome and expose the glass, and the next litho step is to cover certain glass areas and then etch into the glass to a depth that can give the right phase shift. The depth of the opening that gives the phase shift of 180° is defined by the formula $\lambda/2(n-1)$ where λ is the wavelength of light and n is the refractive index of glass.

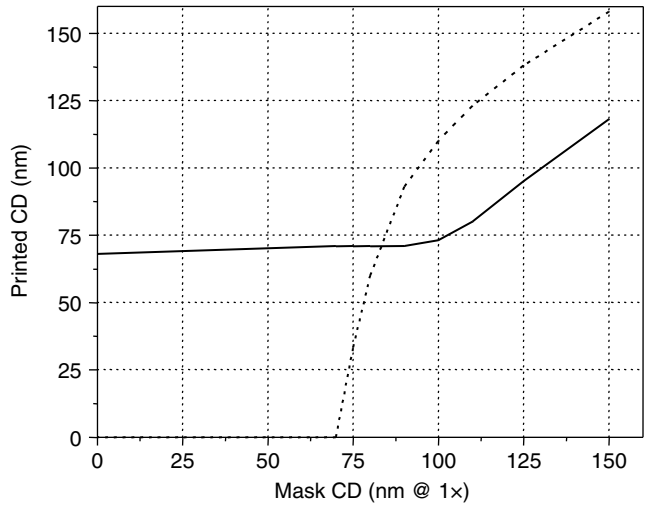


FIGURE 20.8 Printed line width of a dark mask feature between two 150-nm opening as function of feature width. Solid and dotted lines show the effects of with and without phase change. Simulation for 193 nm illumination, 0.75 NA and coherence factor 0.3. (From Maurer, W., and F. Schellenberg, *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

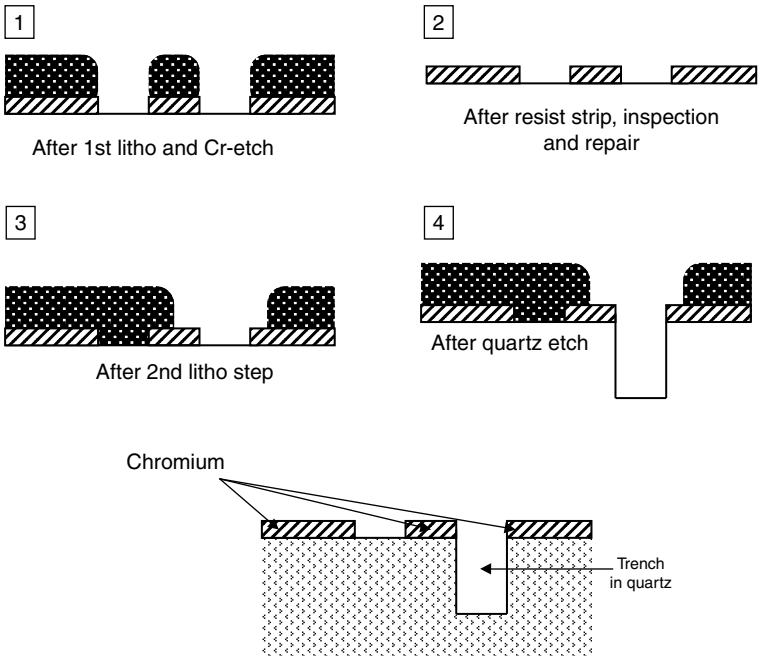


FIGURE 20.9 Manufacturing process of Alt.PSM. (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

One challenge faced in the fabrication of Alt.PSM is the issue of phase balancing [7]. It has been known that two 180° out of phase clear features of the same physical size may not necessarily print with the same size on a wafer. The electromagnetic field transmitted by a trench is quite different from the field transmitted by just an opening between two chromium features Figure 20.10. When the features on the mask are biased so that final outcome on the wafer result in the same size, it may work for one focus setting, but the result will not work for other settings as shown in Figure 20.11. It appears as if the phase difference generated by the mask topology is changed into an "effective phase."

20.2.1.2.3.3 Layout Consideration

Designing of layout where creation of features that are 180° out of phase with an adjacent feature may work in many cases but there still can be cases where such a lay out may not be possible. One such example is the creation of a "T" shaped structure as shown in Figure 20.12. Here it would be impossible to provide a 180° phase difference between the one arm of the structure and its neighbor. This is a good example of what is known as "phase conflict" [8]. Cases like these may require departure from the standard procedure for making Alt.PSM. One strategy could be to arrange phase structure for all the three arms of the structure except for the "cross point" location where the three arms meet. This "cross point" location can then be addressed by a second exposure using another mask known as "trim mask" [9]. This strategy is shown in Figure 20.13. There are other methods to circumvent this type of problems and can be found in the Ref. [2].

20.2.1.2.3.4 Special Cases of Alt PSM

Using phase shift technique it is possible to turn a transparent region into an opaque one.

It is possible because multiple dense phase edges can efficiently diffract light at wide angle. If the spacing between the edges is smaller than (λ/NA) for the stepper, all the transmitted light will be diffracted away from the pupil of the projection lens and none will reach the wafer. The transparent area

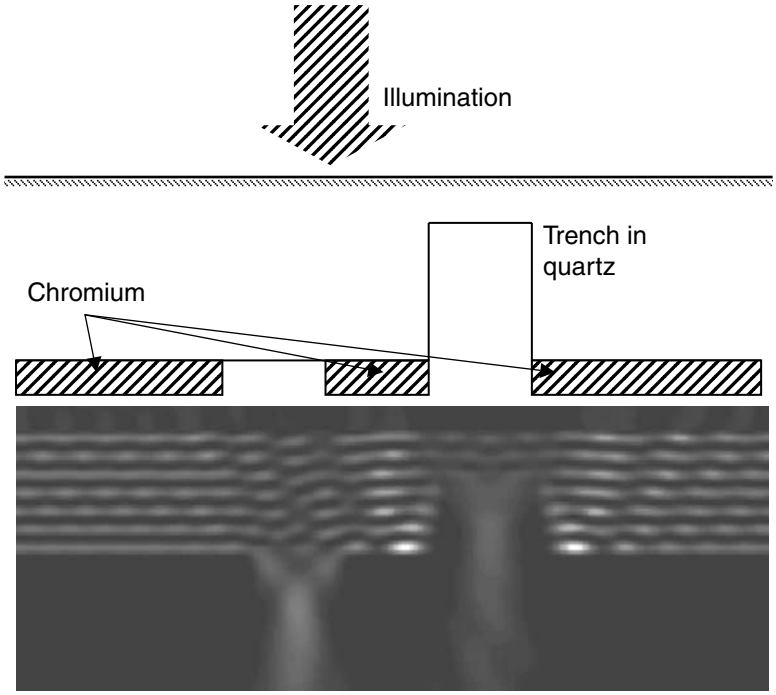


FIGURE 20.10 Two-dimensional intensity plot of light passing the through two openings in an Alt.PSM. Openings are 80 nm each separated by 80 nm chromium, Simulation of 193 nm. (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

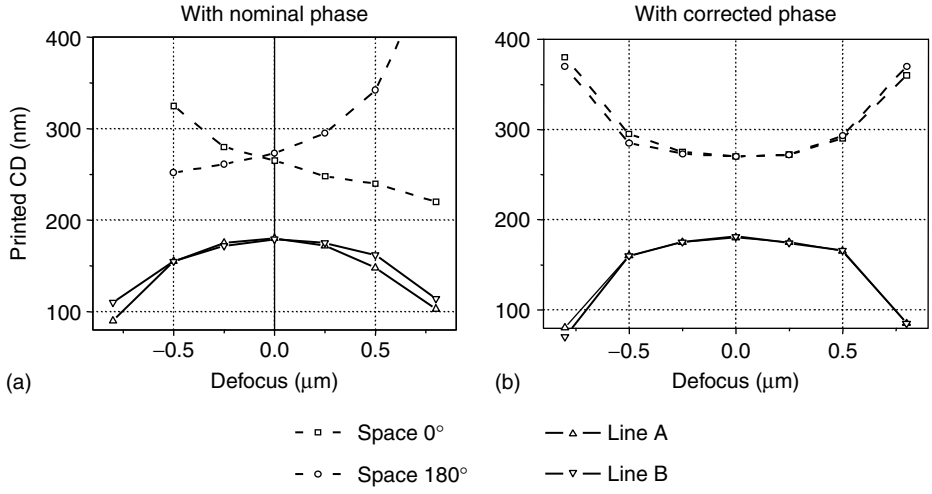


FIGURE 20.11 Case of phase imbalance. The Bossung curve changes for 0 and 180° as shown on the left. Correction made on the right side. (Adapted from Griesinger, U., R. F. Pforr, J. Knobloch, and C. Friedrich, *Proc. SPIE*, 3873, 359–69, 1999 and Maurer, W. and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

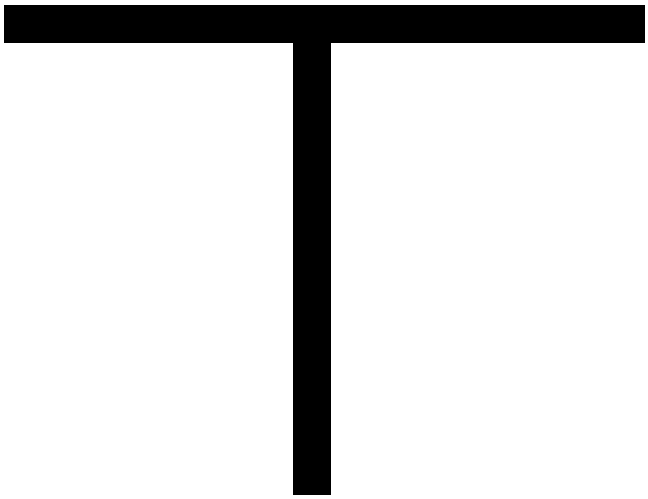


FIGURE 20.12 Example of pattern that can cause phase conflict. (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

will thus print as if it were opaque. This phase approach to create dark regions with dense phase edges is also known as Chromeless PSM [10,11]. Using this approach it is possible to print a feature in form of a single line by having only two phase transition in close proximity and the process has found some application but it does not have an adequate process window using conventional illumination. However when using off axis illumination, it is possible to print a dark line with high contrast and moreover when the distance between the two phase edge increases; the width between the width of the printed line

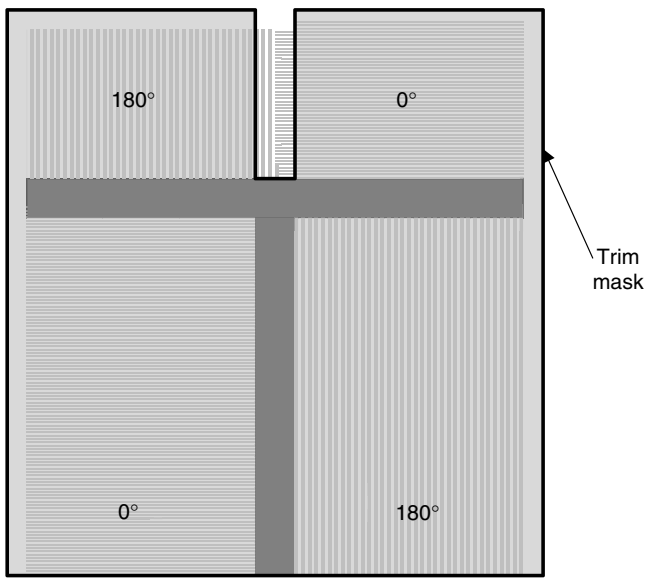


FIGURE 20.13 One solution is the use of Trim masks. (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

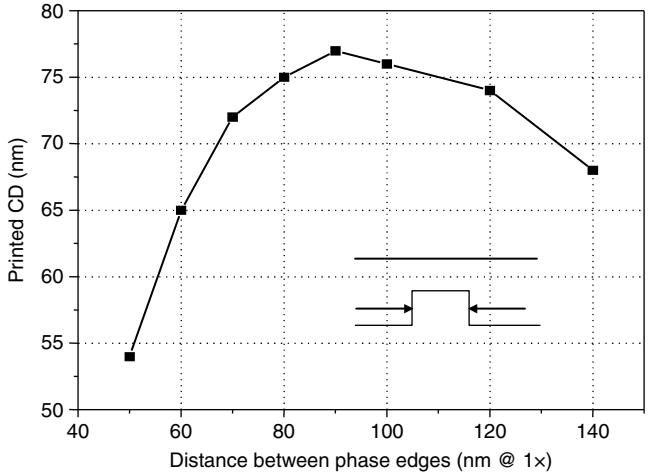


FIGURE 20.14 Printed CD as function of distance between two Cr-less phase edges.

increases slowly providing a lithographic pattern transfer with low MEEF. This behavior is related to what is known as chromeless phase lithography and is shown in Figure 20.14 [2,12]. From the figure it can be seen that as the distance between two phase edge increases beyond a certain point, the printed line will begin to decrease; at this point the contrast also begins to deteriorate [13]. Printing of larger lines would require more innovations like SRAF, or adding chrome in a number of ways as shown in Figure 20.15.

20.2.1.2.3.5 Halftone PSM (HT.PSM)

Principle and Description. Besides the earlier technique for inducing phase shift in mask there is another technique known as HT.PSM which is relatively simple and etching in quartz is not required. In this case

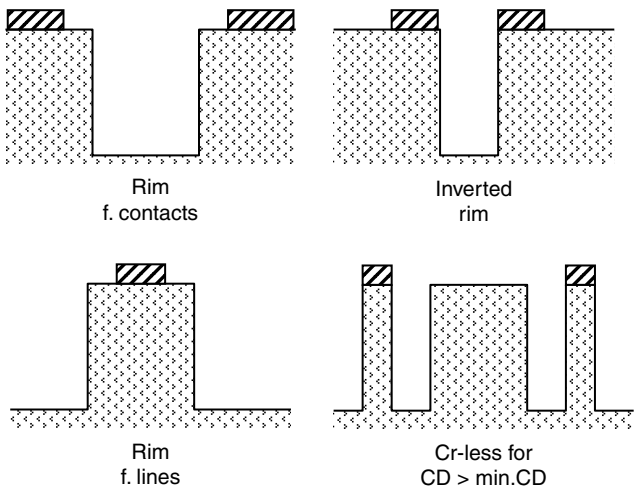


FIGURE 20.15 Various structure of Cr-less mask with off-axis illuminations. (From Maurer, W., and F. Schellenberg, *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

patterns on masks are made of some kind of an attenuating film that, unlike chrome, does allow a small amount of light to pass through but it also goes through a 180° phase change. Because of this attenuating nature of the film these masks are also known as attenuated phase shift mask, or embedded attenuate phase shift mask (EAPSM). The small amount of light that does pass through is not strong enough to expose the resist but, because of its 180° phase difference from the neighboring light, it creates an interference effect to give a zero-intensity at the boundary resulting in a clearly defined border. Principle of this mechanism is explained through Figure 20.16.

20.2.1.2.3.6 Fabrication of HT.PSM

The fabrication process of HT.PSM is similar to that of COG masks. Like transferring pattern from resist to chrome here we transfer pattern from resist to the under laying layer of attenuating absorber film. However there is one issue to be addressed. On wafer we have scribe lines where most of the test patterns and alignment patterns are placed. These are made by overlapping adjacent exposure of the same mask and the feature in the scribe lines have to be delineated in 100% opaque material. This means that the blanks for the HT.PSM material should also have chrome film in addition to the attenuating film. For this reason the chrome film is deposited on top of the attenuating film. Hence the process requires two writing steps. In the first step the chrome film is patterned followed by pattern transfer to the under laying attenuating layer. The next writing step is used to remove chrome from the active area while leaving them in the scribe lines. The second step does not require a very high resolution and overlay specs are not demanding and hence can be done with a less expensive writer in less amount of time. In most cases the attenuating layer is made from ternary compounds such as metal silicides/nitrides where phase or transmission of the material can be adjusted or tuned. Among the ternary compounds, most commonly used materials are molybdenum silicide (MoSi) although MoSiON films are also used [14]. Mixtures of various chrome oxides have also been used for I-line (365 nm) HT.PSM materials [15]. Commercial mask blanks are available with transmissions of 4 and 8% for use with I-line steppers and 6% for DUV (248 and 193 nm) applications. Tuning of the material composition and ratio allows adjustment of optical properties. Mask blanks with transmission values anywhere from 4 to 20% are now available. Moreover, it should also be taken into consideration that the transmission of HT film

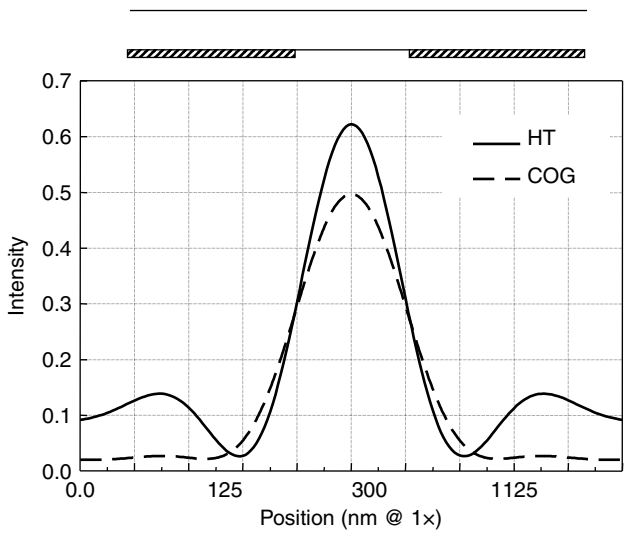


FIGURE 20.16 Intensity profiles for COG and HT trough a 125 nm opening. (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

TABLE 20.2 Typical Transmission Values 193 nm Halftone phase-shift mask (HTPSM) Blank

Wavelength (nm)	Transmission (%)
193	6.0
248	26.1
365	53.2
436	56.8
633	63.3

Source: From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, ed. Rizvi, S., Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.

depends drastically on wavelength as is shown in Table 20.2 for a MoSi HT.PSM material tuned for use with 193 nm steppers [16].

20.2.1.2.4 SRAF and OPC

In the PSMs discussed above alterations in the physical structure of masks were made to induce phase change in the light passing through the masks to compensate for the diffraction effects that are the main cause of pattern degradation and loss of resolutions. These changes were made by etching trenches in the glass substrate and/or introducing new material on the substrate.

The other approach for improving the resolution is to modify the exciting pattern such that the effect of diffraction causing undesirable effects on the wafer are counteracted by the diffraction from the modified features. Moreover the modifications are so small that after undergoing demagnification from the projection optics they vanish and do not print.

A familiar example of such modification is reshaping the corner of a feature in form of very small squares that is done to compensate for corner rounding. These small squares, known as serifs, have been used in the past on the bill-boards and signs in order to make the letters more readable by the observer. Adding serifs on mask features however is not the ideal way to address the problem because they could cause bridging in dense patterns.

There are more effective ways for pattern modification known as SRAF and OPC that forms the subject of the following topics.

20.2.1.2.4.1 Sub Resolution Assist Features

The depths of focus for isolated features are known to be considerably less than for features within a dense pattern. When another feature is placed in the proximity of the isolated feature a desirable effect can be obtained. These additional features are extremely small and are not designed to be resolved and printed on the wafer, and for this reason they are called SRAF. These added isolated features are also known as scattering bars.

Patterns with SRAF have some particular diffractive properties that can improve image fidelity and process window. Making of these SRAFs that are typically 30% smaller than other small features puts special demands on the capabilities of the manufacturing and inspection equipment. Figure 20.17 shows the imaging properties of an aerial image of an isolated line with and without SRAF, using OAI. In this picture an increase in slope at the resist threshold due to SRAF is quite noticeable. Figure 20.18 shows layouts for "before" and "after" the addition of SRAF. Sub-resolution assist features are applied on a variety of mask types that include COG as well as HT.PSM. In fact, an early application of phase shifting used SRAF to increase pattern fidelity and depth of focus [17,18].

20.2.1.2.4.2 Optical Proximity Correction

Basics Principle and Description. A major issue at low k_1 is the non-linearity of the pattern transfer process. Ideally all reduction process should be linear which means that pattern imaging reduces the size of all patterns by exactly the same factor. In reality this is not the case and processes in general have a certain degree of non-linearity. A major contribution to the non-linearity comes from the diffraction of

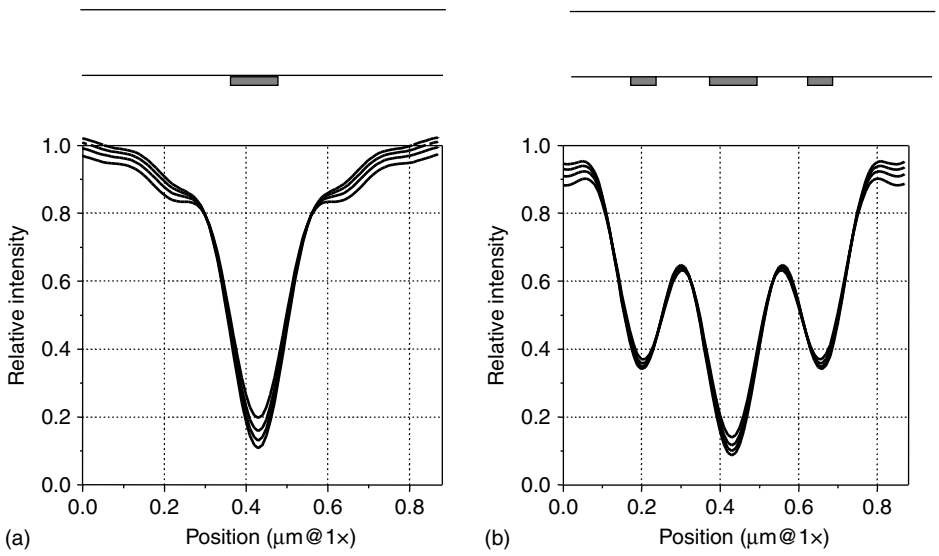


FIGURE 20.17 Simulation without sub-resolution assist features (SRAF) (a) and with SRAF (b). (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

light and is known as optical proximity effect. In order to compensate for these diffraction effects new feature is added to the layout or the layout is modified. These modifications are seen as corrections to the existing layout and are known as OPC.

Initially the corrections used to take into considerations the optical effects only, but soon the effects of resists processing and etch processing began to be included into these considerations. Therefore OPC began to be regarded as an acronym for Optical and Process Correction instead of Optical Proximity Correction [19].

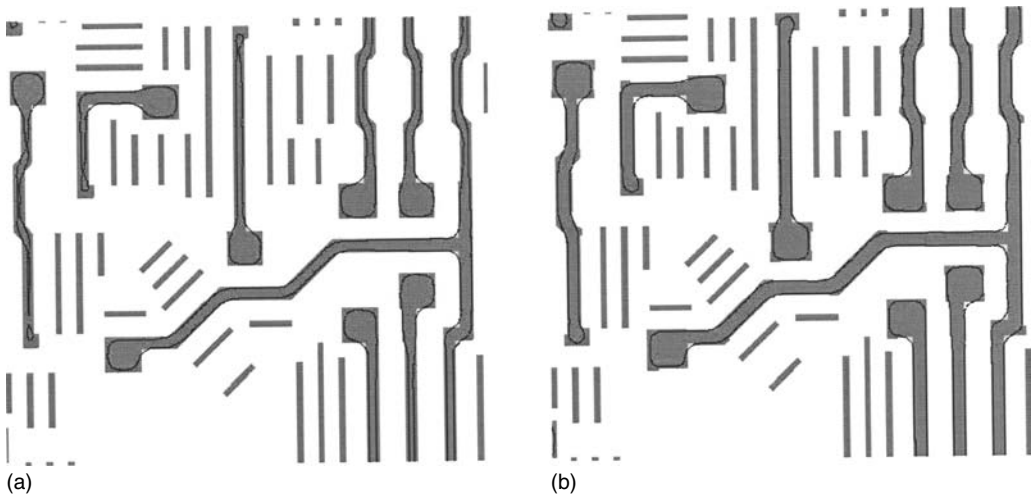


FIGURE 20.18 Layout with SRAF before optical proximity correction (OPC) (a) and after OPC (b). (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

Depending on its type, the designing of OPC is either called rule-based or model-based. Here model-based is the case where behavioral model is used.

In either case the fist step toward designing of OPC is to print a test pattern on a wafer and measure the results. The result is then used to calibrate a description of the process behavior and to describe the non-linearities of the process. This description then becomes the "process-model." This "process-model" can be as simple as a set of a few rules and we call it "rule-based" model. On the other hand, this process-model can be quite complex and may require sophisticated simulation (behavioral description); we call it—"simulation based" model or for brevity we call it "model based." This is not to be confused with "process-based" model.

A simple example of "rule-base" model could be something like saying "nested lines print 15 nm smaller than isolated lines." In "simulation based" model the imaging process is described in some detail. This can be a simple simulation, such as the formation of aerial image using only a few parameters like wavelength, NA, and illuminating conditions, or it could be something more complex that could include a vectorial description of the light propagation in resist, parameters to describe lens aberrations, and various diffusion models for the resist process.

When talking about applying the correction to these models "rule-based" or "simulation or model base" there also can be two approaches to make the corrections.

One type of correction is by inverting the model that can be applied to "rule-base" as well as to "simulation or model base" cases. In case of the above "rule-base" for example, the correction will be to "make nested lines 15 nm larger than the isolated lines."

The other type of correction can consists of an iterative process, where a trial correction is created by moving the mask feature systematically and doing simulation check after each move to see how far the printed image is from the normal image. Both, the trial suggestion and the simulation engine itself must be calibrated against the test pattern measurements. The result of such an approach is illustrated in Figure 20.19.

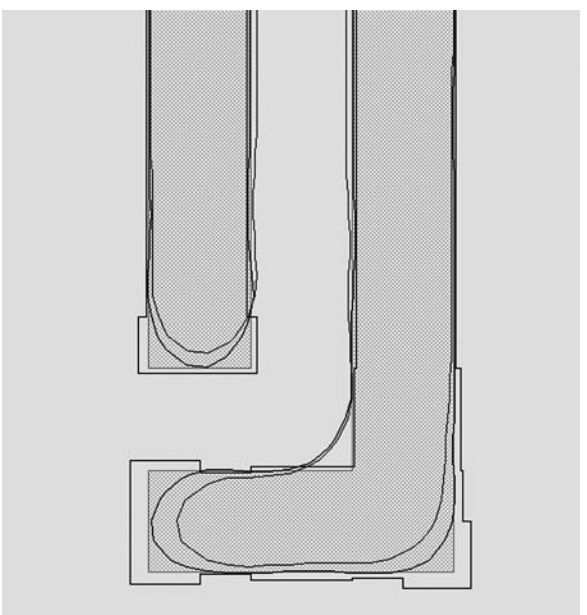


FIGURE 20.19 Optical proximity correction principle. Edges of nominal dark area are broken into segments. (From Maurer, W., and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

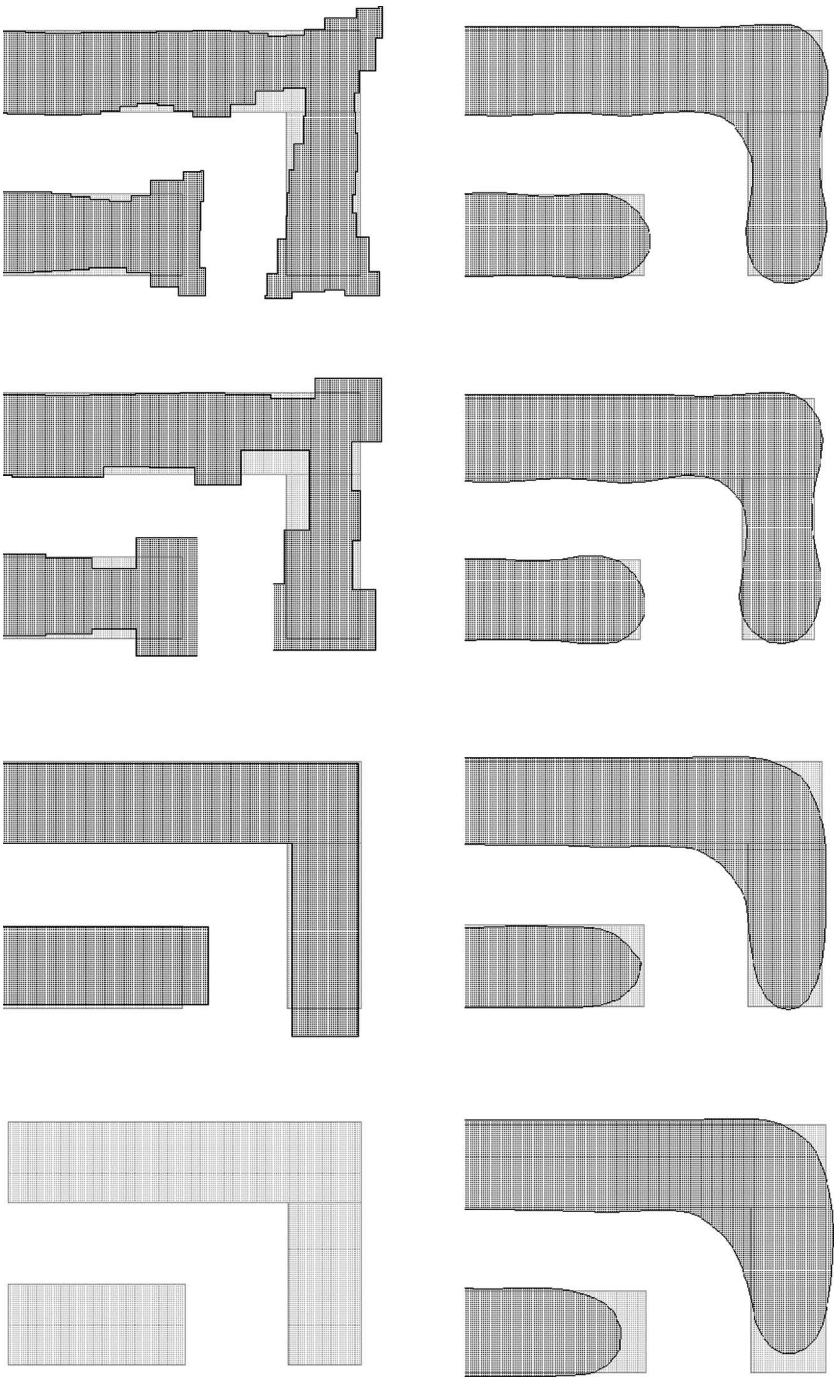


FIGURE 20.20 Different types of OPC fragmentations. (Adapted from Dolainsky, C., and W. Maurer, *Proc. SPIE*, 3501, 774–480, 1997. and Maurer, W. and F. Schellenberg, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 163–89. Boca Raton, FL, 2005.)

Another aspect of OPC, that is particularly important for mask making, is the fragmentation of the mask pattern after OPC. The simple strategy is to simply shift the already existing feature edges of the pattern. However this cannot correct for an abrupt change in proximity of a long line. Therefore OPC tools have evolved over a period of time some sophisticated strategies how to fragment mask patterns. Some such options are shown in Figure 20.20 [2,20].

Due to cost and expense associated with the development and maintenance of highly complex rule set for complex OPC layout there has been a gradual shift from ruled-based OPC to model-based OPC.

20.3 Writing Patterns on Masks

In the earlier section we talked about the structure of various kinds of photomasks. We described the basic structure of photomask with an example of a generic COG mask and we showed how those masks have evolved into today's advanced masks like PSMs, and masks with OPCs and SRAFs. But regardless of what kind of mask we deal with, all masks have some kind of pattern on them. These patterns represent the layout of some circuit that is to be imaged on wafers.

In the following sections we will talk about the techniques involved in making such patterns on masks.

20.3.1 Mask Writers and Data Preparation

Currently there are two types of writing machines used for making patterns on masks. These machines are known as electron-beam writers and laser writers, where depending upon their type they use electron beam or laser to expose the photoresist on the glass plate that later turns into a mask with pattern. Since the purpose of these writings is to draw pattern on mask that represents the layout of a circuit, it is thus necessary that the writing machines receive instructions what kind of patterns are to be drawn on the mask.

These instructions must come from the source where that circuit lay out is designed to start with. However the language in which the circuit-design data is written is not readable by the mask writers and hence the data needs to be translated into machine's language. This task of converting the design data into machine readable data is known as mask data preparation (MDP) and will be addressed in a later section.

20.3.1.1 E-Beam Writer

The birth of electron-beam (e-beam) writing came about in the early 1960s when the scientists at Bell Lab converted a scanning electron microscope (SEM) into an e-beam writer. Soon after, a commonly used polymer polymethyl methacrylate (PMMA) was found to be ideal for an e-beam resist. Interestingly enough PMMA is still used as one of the e-beam resist in the industry.

Basically, an e-beam writer is a machine where an e-beam starting from its source, e.g., LaB6 is made to pass through an assembly of electrostatic lenses and is manipulated by a set of deflectors and beam blankers to expose the resist film on mask. The manipulation of the beam creates the desired pattern on the mask. The spot size of the beam is very small which determines how small a feature can be resolved on the film. Today's e-beam can be focused into nanometer diameter and can be steered on the resist surface with great precisions.

The role of e-beam lithography (EBL) in semiconductor manufacturing can be illustrated by Figure 20.21 [21]. The initial layout data is generated using a computer aided design (CAD) system. The CAD data is then converted to EBL data and then the EBL system produces the final pattern. Primarily the e-beam writers are used for making patterns on masks but in some special cases they are also used for writing directly on wafers. An architectural example of an EBL system is shown in Figure 20.22.

20.3.1.1.1 EB Source and Optics

The electrons used for writing the pattern are extracted from their source mounted near the top of the column where the column is furnished with a set of electrostatic lenses and deflectors. The electrons are

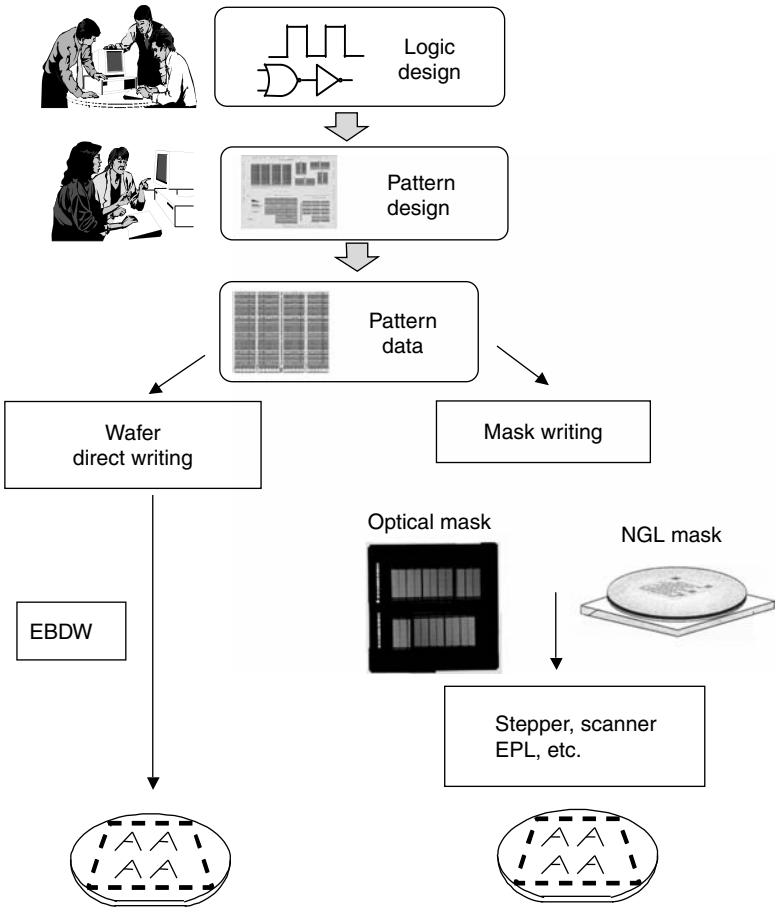


FIGURE 20.21 E-beam lithography for wafers and masks. (From Saitou, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 59–98. Boca Raton, FL, 2005.)

accelerated and made to pass through the entire column where they are focused to form a spot on the mask coated with photoresist. The deflectors in the column are used to move the spot in order to write the desired pattern of the mask. The instruction to move the spot in accordance with a pattern to be written comes for a set of software programs that is driven by the design data.

20.3.1.1.2 Scanning Mechanism and Stage Movement

Electron beam lithography systems can be classified under various categories depending upon things such as their scanning methods, beam shapes, or stage moving strategy, etc. As we will see to a great extent all these characteristics are also interrelated but since the classification in terms of scanning mode is most commonly referred to, we will start our discussion from this type of classification.

These classifications in terms of scanning modes are: (a) raster scanning mode and (b) vector scanning mode. Schematics of these modes are shown in Figure 20.23.

20.3.1.1.2.1 Raster Scan System

In raster scanning mode a deflector mounted inside the column scans the beam over an entire field. It is done in such a way that the beam is "blanked off" in the non-pattern region and is "un-blanked" in the

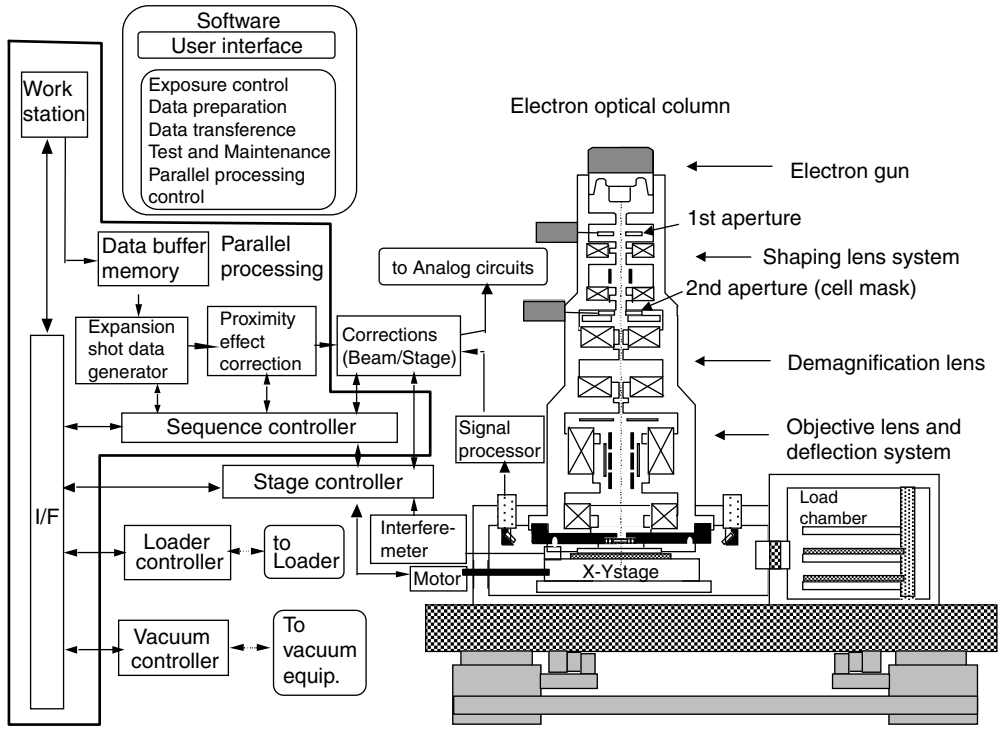


FIGURE 20.22 Example of an e-beam set up. (From Saitou, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 59–98. Boca Raton, FL, 2005.)

pattern region. The system structure for raster scan is relatively simple because it does not depend upon pattern complexity and entire region of the substrate surface is uniformly scanned. The raster scanning method is usually combined with the continuously moving stage. In raster scanning, the field size to be scanned is small and thus it puts less constraint on the optics.

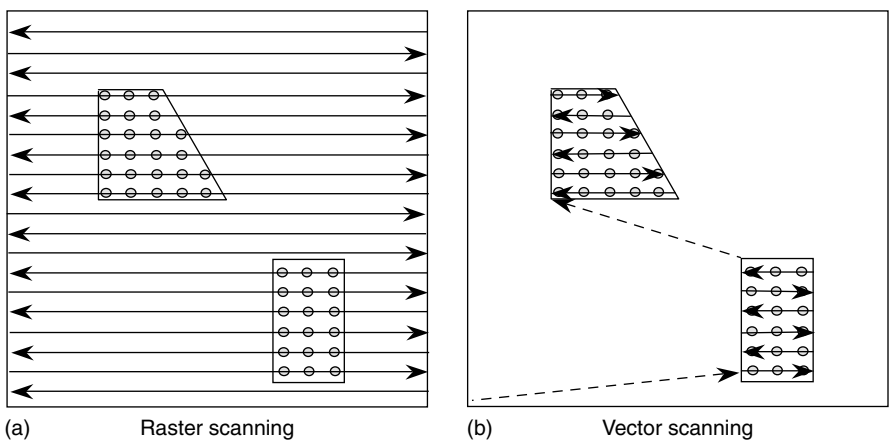


FIGURE 20.23 Example raster and vector scan. (From Saitou, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 59–98. Boca Raton, FL, 2005.)

20.3.1.1.2.2 Vector Scan System

In vector scanning mode the deflector scans the beam only in the pattern region. In this case the scanning time can be shorter because unlike the raster scan the entire area does not need to be scanned. Also, in this case field size can be larger and give higher throughput. But larger field size could lead to distortion problem which must be controlled and hence the system could be subjected to more constraints.

20.3.1.1.2.3 Stage Movement

Due to the limits caused by possible optical aberration and distortion, the scanning field size of deflector is kept below a few mm square. In order to cover a larger field size the stage movement is essential and must be very precise.

There are two modes of stage movement as illustrated in the lower half of Figure 20.24. The first one is the step-and-repeat (S&R) method. In this case the machine does the writing inside a field, when the writing is completed the stage moves to the next position and the process is repeated. The other method is where the stage moves continuously while the writing is carried out. In this method, the deflection aberration tends to be less serious than S&R method, because the field size here can be made much smaller than in the S&R method. Considering the time factor, there are lot of dead times in the stage movement in the first case but in the second case the scale is in a continuous motion and the system gives considerable higher throughput. In the second case however, the field stitching error problem becomes serious because the number of field stitching increases with decreasing field size.

20.3.1.1.3 Beams and Their Impact on Imaging

20.3.1.1.3.1 Two Types of Beams

There are also options what kind of beam can be used for writing the pattern on mask. The beam can be a "point-beam" or a "shaped beam."

In one system the image of the source is focused to a sharp round spot on the substrate.

This spot has a Gaussian distribution as regards to its intensity and is referred as a point beam. Figure 20.25 (a) shows a schematic of the optics that produces such a spot. A drawback with points shape systems is that their throughputs are relatively low but give better resolution than obtained from the other types.

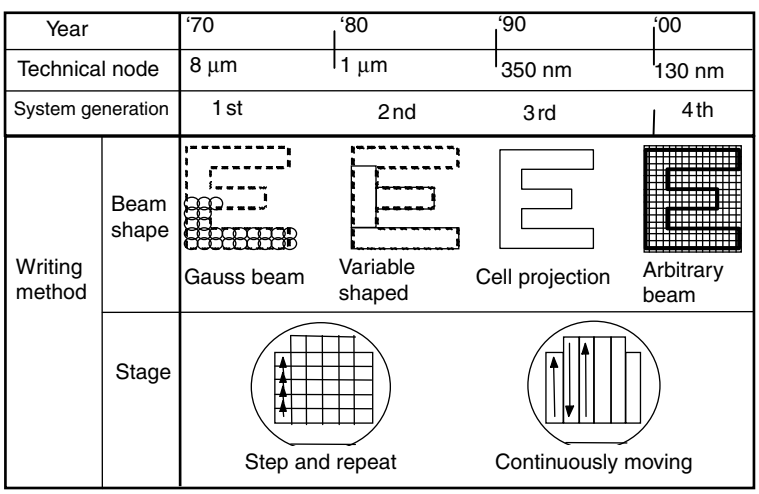


FIGURE 20.24 Evolution in e-beam scanning. (From Saitou, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 59–98. Boca Raton, FL, 2005.)

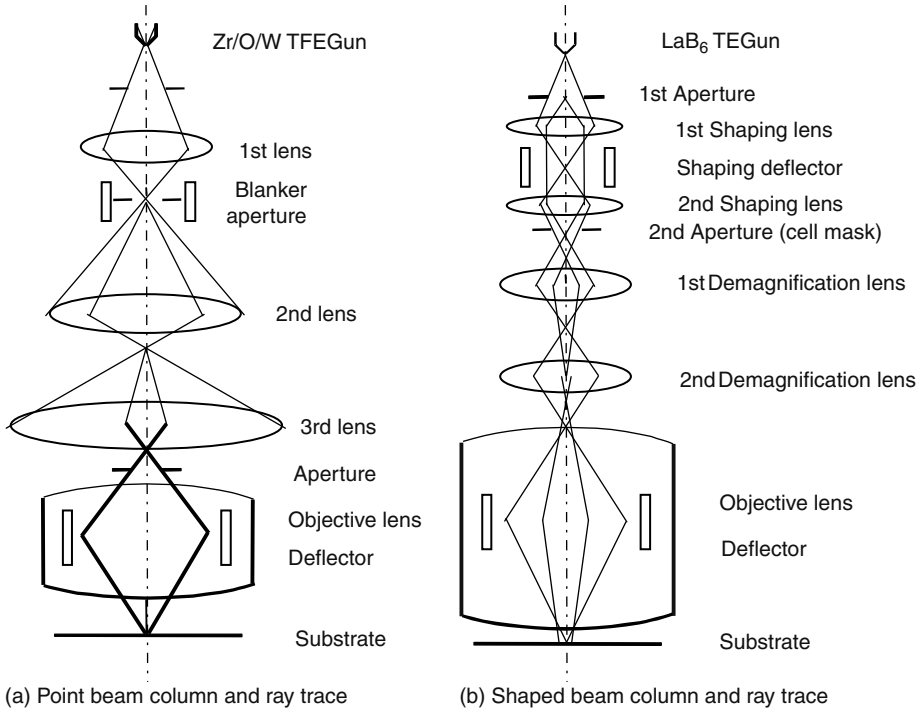


FIGURE 20.25 Optics for point beam and shaped beam. (From Saitou, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 59–98. Boca Raton, FL, 2005.)

The other type of beams that are commonly used is "shaped" beam. Here the shape of the beam can be varied during the writing of a pattern and hence it is commonly known as variable shaped beam (VSB). An example of VSB can be seen in Figure 20.24 where the figure shows how a character "E" can be delineated by four VSB shots in contrast to a large number of shots that would be necessary for a point beam to produce.

Formation of VSB can be seen form the schematic in Figure 20.25b. Here a rectangular beam with variable size is created by passing it through two square apertures.

A beam spot with a desired shape can then be projected on to a substrate on a shot by shot basis. Unlike the case of "point" beam the source image of the electron gun is not focused to the substrate in this case.

There is yet another category known as cell projection (CP) that can be regarded as a sub-category of "shaped" beam like VSB is. In this case several unit cell patterns are defined in a second stencil mask. The required pattern forms the stencil mask which is illuminated by the square from the first mask and projected on the wafer. The process can then be repeated in form of an array as shown in Figure 20.26. The cells are made of a thin silicon stencil structure using conventional wafer fabrication process [22].

In VSB as well as in CPs the patterns are de-magnified by 1/25–1/100 size through two de-magnifying lenses and an objective lens system.

20.3.1.1.3.2 Resolution, Accuracy, and Throughput

The three key parameters to characterize an e-beam system are resolution, accuracy, and throughput. One system, depending upon its application may be strong in one area while the other system may be strong in some other area. Figure 20.27 is a three-dimensional representation of three separate e-beam systems where each system shows its strength in one of the three parameters. The first system (a) has a high resolution of the order of 10 nm; the second system (b) is strong in accuracy; while the third system

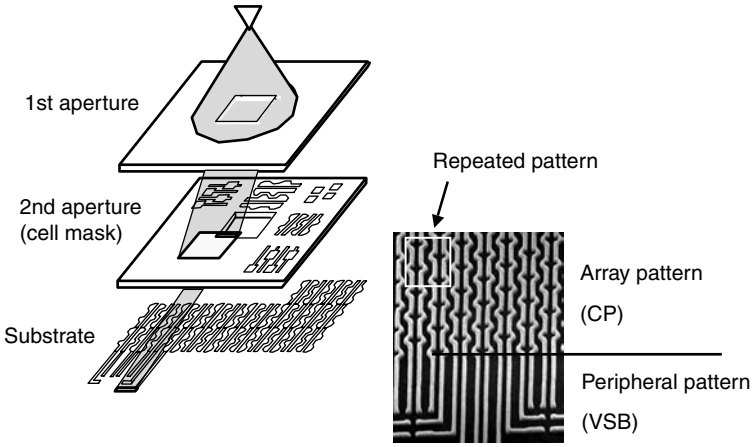


FIGURE 20.26 Formation of shaped beam. (From Saitou, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 59–98. Boca Raton, FL, 2005.)

(c) shows a throughput that supercedes the other two. Strengths in each of these parameters can be of value to a specific task. For example the first system with strength in resolution can be very valuable for R/D, whereas the second system with high image placemen accuracy would be more useful in mask fabrication. The third type with high throughput can definitely be of a great value in "direct write on wafer" programs, where throughput is of prime importance. The volume of the solid triangle defined by these parameters can be regarded as the performance capacity of these systems. It is noticeable that the volumes of these three solid triangles are almost equal to each other. It can therefore be surmised that the performance capacity of a system is a measure of the technology level of the era or of the system manufacturing company. During the last three decades this performance capacity has grown over one thousand times and is still growing.

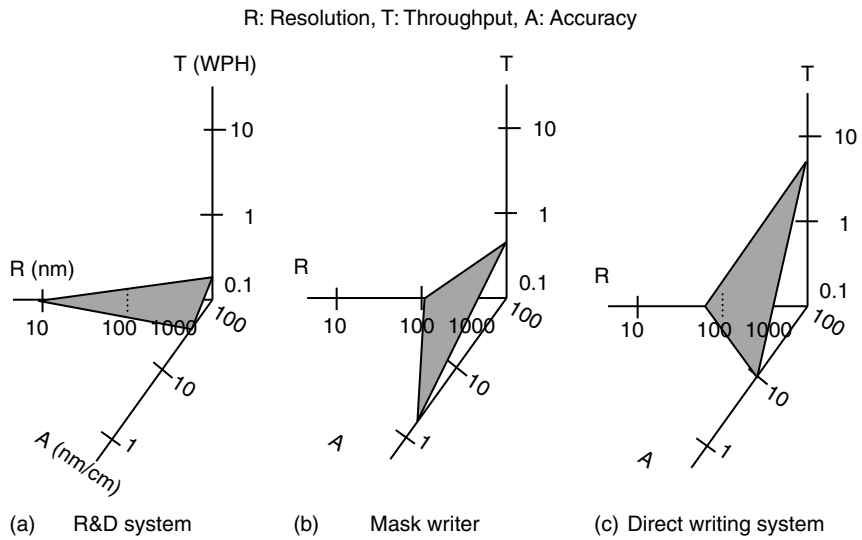


FIGURE 20.27 Performance of different types of e-beam systems. (From Saitou, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 59–98. Boca Raton, FL, 2005.)

20.3.1.1.3.3 Acceleration Voltage

Acceleration voltage is the driving force of any e-bam system, since it is this force that is responsible for extracting the electrons from their source and moving them on their path to reach the plate. The acceleration voltage V should be selected carefully. The acceleration voltage of the system relates to the throughput because resist sensitivity and current density depend on the acceleration voltage. The acceleration voltage also relates to the resolution, the CD accuracy, and the IP accuracy of the e-beam system due to Coulomb effect and proximity effect.

20.3.1.1.3.4 Space-Charge Effect

Electrons falling on their target may give rise to a space charge effect resulting in deterioration of the beam edge sharpness. Due to coulomb effect from the space charge the sharpness of the beam may transform into beam blur which is proportional to $IL/V^{3/2}$, where I is the beam current, and L is the optical path length [23]. A larger beam current can increase the throughput but it will do so at the cost of resolution. In a VSB system the current changes shot to shot during pattern writing and therefore the beam resolution can be different for each shot. As the fast refocusing shot to shot makes the e-beam system so complicated, most VSB systems adopt high acceleration voltage and limit the maximum current depending on the accuracy. However, the CP system is able to have the refocusing function for each cell pattern, because the time for refocusing is not so short for each cell.

Acceleration voltage lower than 10 kV causes charging-up in resists which deteriorates the position accuracy. It needs thin resist process but thin resist has defect problems. At the 100 nm node, the most suitable acceleration voltage might be 50 kV as adopted in almost commercially available e-beam mask writers.

20.3.1.1.4 Resist in E-Beam

20.3.1.1.4.1 Resist Sensitivity and Current Density

Resist sensitivity has been known to be inversely proportional to the accelerating voltage hence low accelerating would be preferable where high throughput is required. However, due to recent development of highly sensitive and chemically amplified resist (CAR) higher accelerating voltages are also being used for increased throughputs.

20.3.1.1.4.2 Proximity Effect

When the electrons reach their target (e.g., photoresist in this case) they undergo all sorts of scattering depending on the accelerating voltage of the system and on the material of course. An obvious effect of such scattering is the broadening of the beam. The scattering takes place while the electrons are moving forward on their path and during the process many electrons are back-scattered and cause undesirable exposure to resist that were not meant to be. The scattering phenomenon causes the proximity effect in e-beam pattern writing. This phenomenon can gravely affect the patterns when they are close together.

The forward and backward scatterings mentioned here, are the result of the deposited energy distribution as expressed by the following Equation [21] which is a double Gaussian expression. Here the first term represents the effect of forward scattering in the resist and the second term is the backward scattering from the substrate.

$$f(r) = \frac{1}{1+\eta}\left\{\frac{1}{\beta_f^2}\exp\left(-\frac{r^2}{\beta_f^2}\right) + \frac{\eta}{\beta_b^2}\exp\left(-\frac{r^2}{\beta_b^2}\right)\right\}$$

Here r is the distance from the incident point, β_f is the forward-scattering range, β_b is the back-scattering range, and η is the ratio of the backward-scattering energy to the forward-scattering energy.

The degree of beam broadening has also been studied by applying Monte Carlo simulation as shown in Figure 20.28 [21,22,24]. In this simulation, one hundred electrons were impinged perpendicularly to one point of the surface and the trajectories of electrons were projected in x–y plane. It can be seen from the figure that resolution in EBL can be adversely affected by the scattering of the beam.

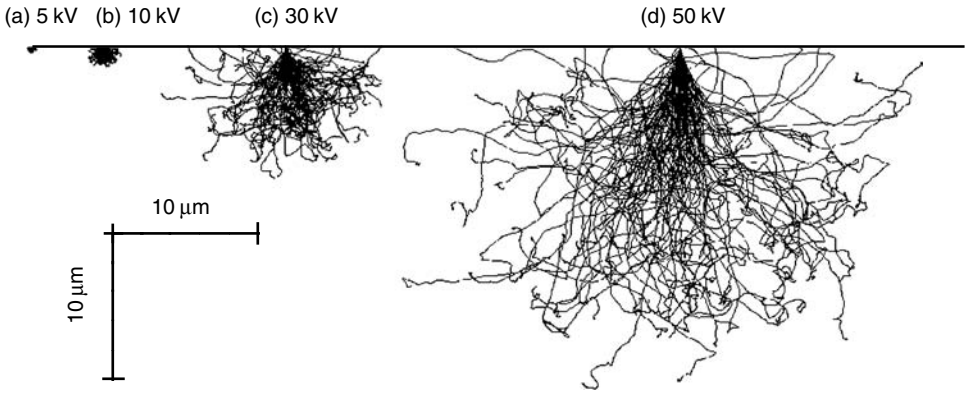


FIGURE 20.28 Simulation of e-beam trajectory. (From Saitou, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 59–98. Boca Raton, FL, 2005.)

The penetration depth of electron is known to increase with the increase in the accelerating voltage and is proportional to $V^{3/2}$. For example, for acceleration voltage 5, 10, 30, 50 kV the penetration depth happens to be 0.7, 2, 10, 20 μm. At the acceleration voltage 10 kV or below the lateral spread of the electrons at the surface is almost the same as the penetration depth. At higher than 30 kV, the lateral spread becomes very small. The thickness of resist on mask is about 0.3–0.4 μm where the scattering situation is the same as described here.

Typically, the scattering phenomena can be classified as two types. One is forward scattering effect and the other is backward scattering effect as shown in Figure 20.29. It is apparent from the picture that the effect on the resist from the forward scattering is smaller than that form the backward one. The forward case gives rise to a intra-proximity effect which remains confined within a pattern. Moreover, the pattern fidelity to the beam shape is generally better using a high acceleration voltage for sub-micron isolated pattern. This can be seen from Figure 20.29. The backward scattering on the other hand gives rise to a

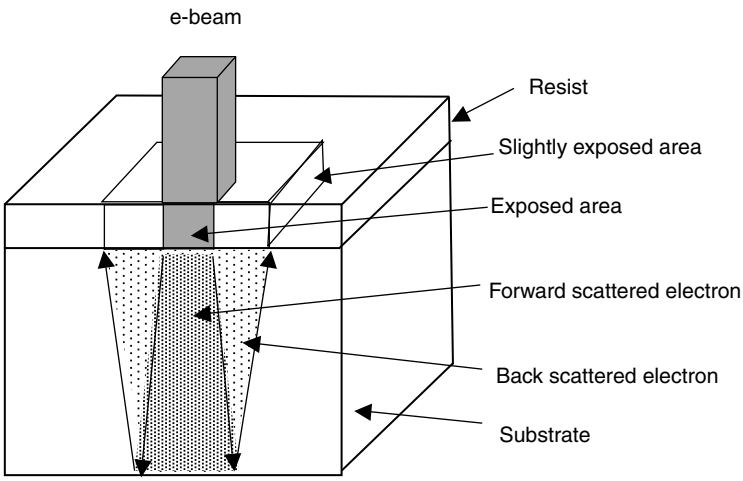


FIGURE 20.29 Forward and backward scattering on proximity effect. (From Saitou, N., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 59–98. Boca Raton, FL, 2005.)

inter-proximity effect between patterns shown in the same figure. Acceleration voltage lower than 10 kV causes charging-up in resists which deteriorates the position accuracy. It needs thin resist process but thin resist has defect problems. At the 100 nm node, the most suitable acceleration voltage might be 50 kV as adopted in almost commercially available e-beam mask writers.

Proximity Effect Correction (PEC). PEC is not to be confused with OPC discussed in an earlier section. Although both are corrections related to proximity of a feature, the basis of their origin differs. While in the current case features size can be affected by the scattering of electrons in the resist whereas in the other case the features are affected by the diffraction of light which is purely optical in nature. There are a number of schemes for PEC as discussed in Reference [21].

20.3.1.2 Laser Writer

Laser mask writers are another option for drawing patterns on photomasks where very high resolutions are not critical. An attractive feature of laser writers is that they are significantly faster than the e-beam writers, and moreover these machines donot require vacuum environment for their operation as the e-beam writes do.

Currently there are two types of laser writers used by the industry. One type falls under the class of raster-scan tools, while the other kind employs spatial light modulators and known as SLM tools.

The earliest type of laser writers, also known as PGs, used to expose the photomask by scanning a single laser beam on the mask surface. These systems were slow for today's standards. A dilemma faced with such a system has been that in order to achieve a high resolution, the beam spot must be made smaller and smaller, but the drawback with this approach is low throughput. The way to circumvent this problem was to use multiple beams without having to increase the size of the spot. This is the approach many laser writers employ today. The first commercial system employing this raster scan technology was CORE™ from Etec Systems Inc. an Applied materials Company [25].

The other kind of writer evolved from the concept of stepper and scanner where the image of a reticle is projected on a wafer. Here, instead of reproducing the image of a reticle pattern on a wafer, a dynamic mask in form of a SLM is employed, and is imaged onto a photomask.

This architecture involves a printing strategy where the reticle pattern is replaced by a pattern that is made up of several individual "stamps" and where these stamps can be changed dynamically. Different technologies have been developed to make these dynamical changes in the stamps.

The first production system for writing photomasks with dynamic SLM imagining was the Sigma product family from Micronic Laser Systems. The Sigma systems use a SLM with pivot micro-mirrors operating in diffraction mode developed by the Fraunhofer IMS and IPMS Institutes.

20.3.1.2.1 Laser Writers Using Raster Scan

The use of raster scan in laser writing has been around since the inception of laser writing technology, and is seen as a matured technology for mask writing. For example the ALTA™ series from Etec Systems, Inc. and the Omega™ series from Micronic Laser Systems use this approach [26,27].

The raster scan PG writes the pattern with one or several laser beams of an approximate Gaussian shape. The beams scan over the mask surface while being amplitude modulated according to pattern data. The writing time for a pattern is essentially proportional to the size of the pattern area and not to the complexity or number of features on the pattern.

Two such raster scan systems are shown in Figure 20.30 and Figure 20.31 [28]. These systems include a laser, a beam splitter, a modulator, either a rotating polygon or an acousto-optic deflector creating a crosswise scanning motion of the beams, a reduction lens, and an X–Y stage. The beam splitter divides the light from the laser into several beams for increased capacity for writing different parts of the pattern in parallel. Each beam is individually amplitude modulated in the acousto-optic modulator. The modulation is controlled by input from the data path. The area to be exposed is divided into stripes of equal width. The width of a stripe, or a scan stripe, is typically a few hundred μm. During exposure the X–Y stage move the photomask at a constant speed along the stripe. At the same time the focused spots

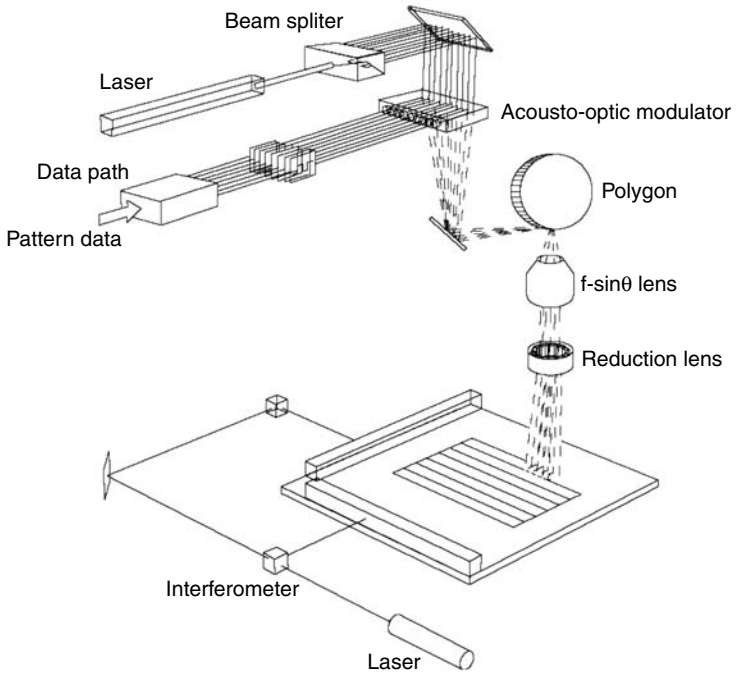


FIGURE 20.30 The ALTA raster scan tool from etec Systems, Inc., an Applied Materials Company. (From Rydberg, C., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 99–132. Boca Raton, FL, 2005.)

scan in a direction perpendicular to the movement of the X–Y stage. A rotating polygonal mirror (typically with 24 facets) or an AOD creates the crosswise scanning motion of the spot Figure 20.32.

The spot is moved continuously in the scan direction and discrete scans are added in the stripe direction. After reaching the end of a stripe and before the start of the next stripe the X–Y stage moves one increment in the direction of the scan and return to the start position in the stripe direction. Due to the small speed vector component from the movement of the photomask, a scan, placed perpendicular to the X–Y stage motion, will not end up perpendicular on the photomask. This small angular deviation from an orthogonal behavior can be compensated for by a slightly tilted scan line, denoted as the azimuth angle.

20.3.1.2.2 Laser Writers Using SLM

Laser Writers of the second kind are based on the same principle as that of the conventional wafer stepper/scanners. In steppers/scanners it is the pattern on a "reticle" that is imaged on wafer, here in this case the "reticle" is replaced by "mask-like device with programmable patterns" that is to be imaged on a reticle. This device is named as a "programmable mask" although this is not a mask in the accepted meaning of the word.

The concept of a programmable mask can be visualized in terms of an array of pixels of light that can be individually turned on or off to create a desired pattern of light that can then be imaged on reticle. One such programmable mask is known as SLM.

As mentioned earlier, a programmable mask like this can be seen as a dynamic mask where its pattern can be created or changed in real time while in use. This is in contrast to the conventional mask where the pattern on it is already written which can be called as a static mask.

An SLM is a device that can control a light field through modulation of the amplitude, phase, or polarization. Micro-mechanical SLMs are quite suitable for such a task. An example of an

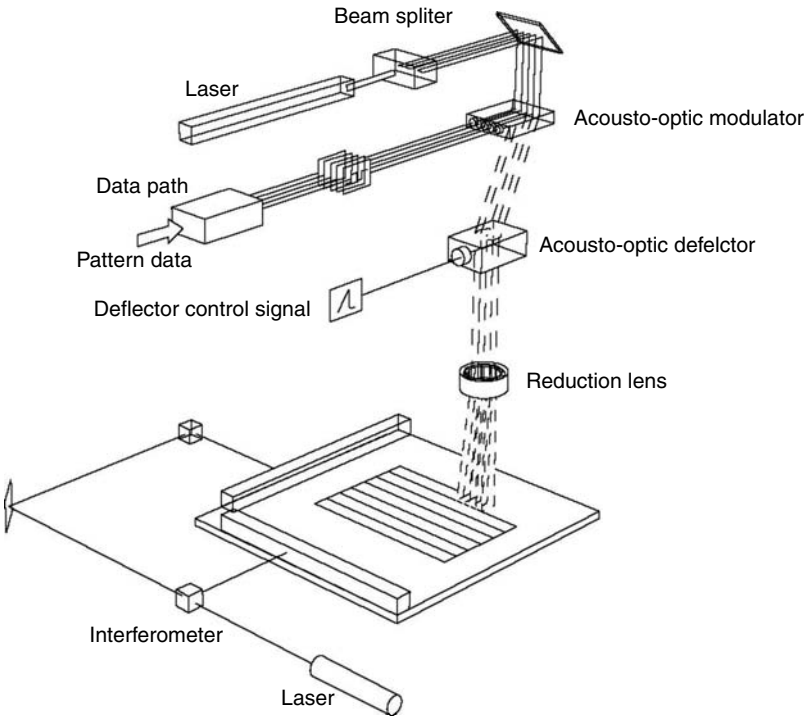


FIGURE 20.31 The Omega raster scan from Micronic laser systems AB. (From Rydberg, C., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 99–132. Boca Raton, FL, 2005.)

amplitude-modulating SLM is known as digital mirror device (DMD) developed by Texas Instruments. In this device the tilting mirrors deflect light away from the aperture of the projection optics. Another interesting group of SLMs is the phase-shifting SLMs where interference effects are used. The main idea here is that the phase of the illuminating field is modulated in cells covering an area. Either interference between the cells or interference within the cell itself cancels the light in a certain direction.

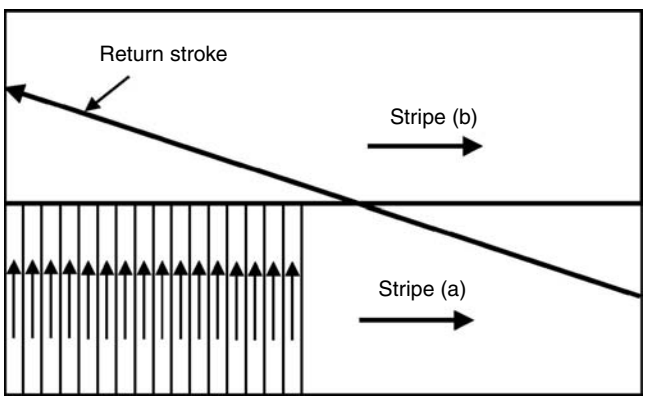


FIGURE 20.32 Raster scan writing principle. (From Rydberg, C., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 99–132. Boca Raton, FL, 2005.)

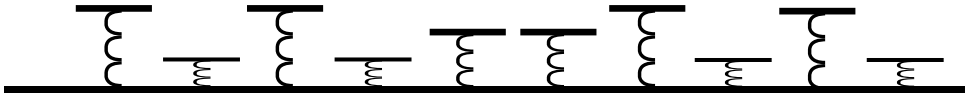


FIGURE 20.33 SLM piston mirror type. (From Rydberg, C., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 99–132. Boca Raton, FL, 2005.)

Here we give one-dimensional description of two SLM devices where both are phase related. These are piston micro-mirrors and pivot micro-mirrors. Figure 20.33 and Figure 20.34.

In the piston arrangement each mirror in a two-dimensional matrix can, by some mechanism, be set at an individual height. With this arrangement the average phase change over each individual mirror can be controlled. In the pivot type the mirror pivot around a central axis. Here the phase remains in a constant arrangement for all pivot angles and displacing the tip of a micro-mirror by quarter of a wavelength gives a liner phase difference of 0°–180° over the extent of the mirror, that is the extinction of the field.

20.3.1.2.2.1 Architecture

The current Sigma family made by Micronic Laser Systems employs the pivot mirror technology on some of its machines shown in Figure 20.35 [28,29].

For example, in one configuration (Sigma 7300™) the SLM has 2048×512 micro-mirrors that are 16 μm×16 μm each and have a projected image on the photomask of 80 nm×80 nm. The device works in a diffraction mode and needs to deflect the mirrors by only a quarter of the wavelength (62 nm at 248 nm) to go from the fully "on" state to the fully "off" state. To create a fine address grid the mirrors are driven to "on," "off" and 63 intermediate values. A pulsed excimer laser is used and the SLM is illuminated with a partly coherent illumination.

The pattern is stitched together from millions of images from the SLM chip. Flashing and stitching proceed at a rate of 2000 stamps per second. While writing the X–Y stage moves continuously in the stripe direction Figure 20.36.

The pattern data is loaded for programmable mask and exposure are made according to the program. When the stage passes over the correct position, the laser is pulsed and projects the content on the programmable mask onto the photomask. The pattern is divided up into several stripes of equal width. Stamps along the stripe direction expose each stripe while the X–Y stage moves continuously. Both adjacent stamps and adjacent stripes are printed with a small overlap to ensure pattern quality at the boundaries. Because the imaging process within a stamp is partially coherent, but the stamps are incoherent to each other, even an image printed with perfect placement needs to use blended overlaps to "dilute" the edge discontinuity. The latest in the Sigma family from Micronic Laser Systems is Sigma 7500. This machine combines two writing modes in a single platform. The standard writing mode delivers almost twice the productivity of Sigma 7300 and with an extended writing mode it provides a significant improvement in pattern accuracy which amounts to 12 nm 3-sigma global registration accuracy.

20.3.1.2.2.2 Fabrication of SLM Chip

The micro-mirror SLM module is made using micro electrical mechanical systems (MEMS) and complementary metal-oxide-silicon (CMOS) processes. First an electronic chip is created in a standard CMOS process. On top of the electronic chip a matrix of individual micro-mirrors with flexing hinges are created by use of MEMS processes. The micro-mirrors, hinge structure, and support posts are formed in an aluminum alloy.

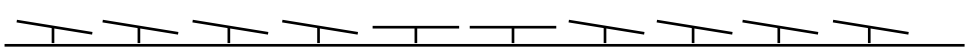


FIGURE 20.34 SLM pivot mirror type. (From Rydberg, C., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 99–132. Boca Raton, FL, 2005.)

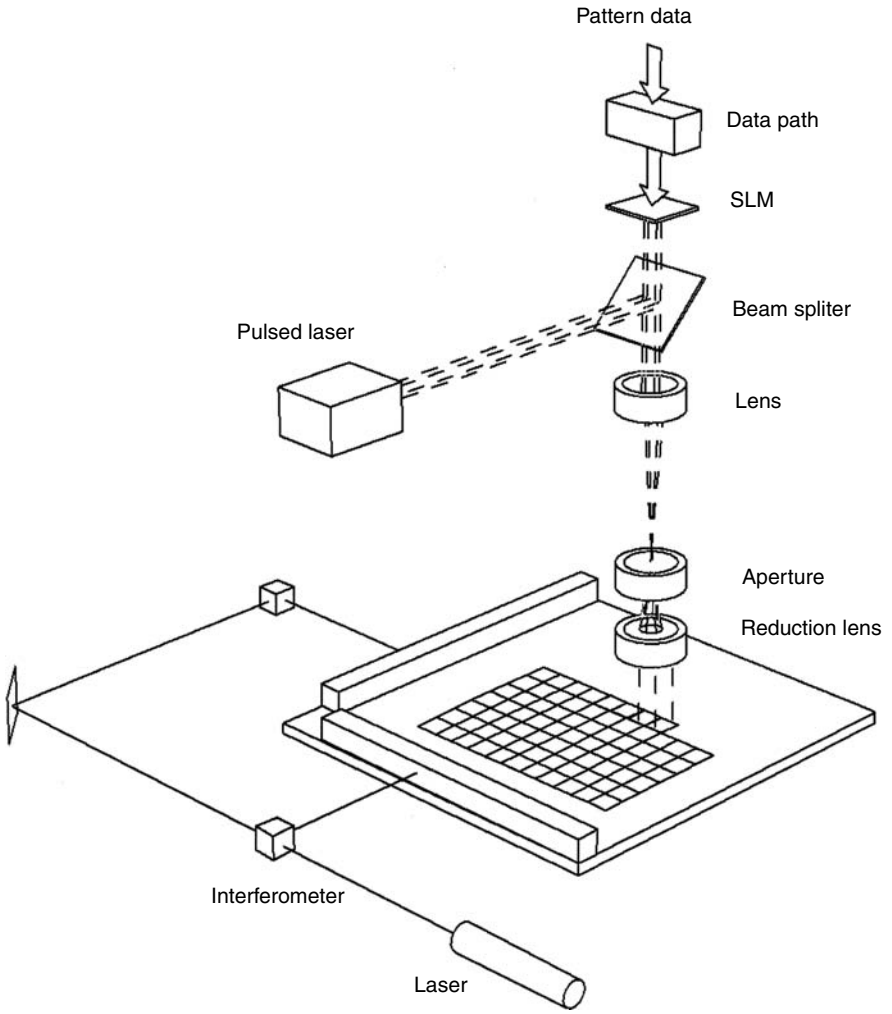


FIGURE 20.35 The sigma SLM tool from micronic laser system AB. (From Rydberg, C., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 99–132. Boca Raton, FL, 2005.)

Underneath the mirror matrix is an addressing system for each individual mirror. Each mirror has a transistor and is addressable in a matrix structure, similar to a TFT display. An electrostatic force is created that cause the mirror to be slightly deflected. The pivot angle applied to each micro-mirror is in the range of a few milliradians.

20.3.1.3 Data Preparation

All work related to the fabrication of mask starts from some kind of a design layout of a circuit which is then transferred onto a photomask in the form of patterns representing the original layout. This lay out is initially generated by a system known as CAD, but the data that defines this layout is not readable by the machine that writes patterns on a mask. CAD data needs to be translated into machine's language and the process is known as mask data preparation or simply MDP. The reason is that the design layout output format is usually an interchange format rather than a machine specific format. Here objective of the MDP is to convert this interchange format to a machine specific format. There are different types of mask writing tools and most of these tools have their own formats.

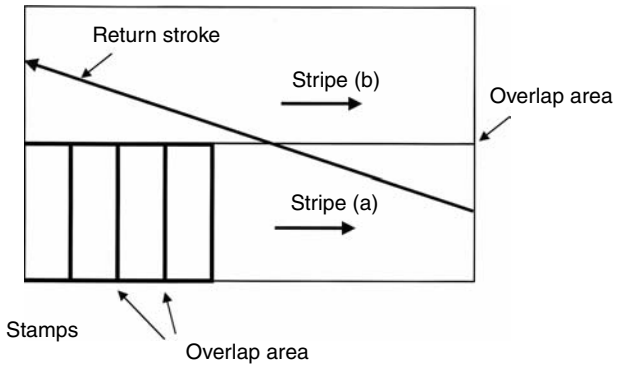


FIGURE 20.36 Sigma SLM writing principle. (From Rydberg, C., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 99–132. Boca Raton, FL, 2005.)

Besides the design data, there are other kinds of data that also need to be written on the mask. There can be test patterns, alignment patterns, and bar codes on the scribe lines, which are not part of the main circuit, but are necessary for monitoring the process.

Hence, apart from just the conversion of the design data, these patterns need to be created for these scribe lines. This type of synthesis of mask pattern is known as "frame-generation" and constitutes a "jobdeck."

Before MDP the data goes through a design rule check (DRC) followed by a layout vs. schematic (LVS) run. These steps ensure verification, measurements, and the manufacturability of the entire set that may consist of as many levels as 30 or even more. Figure 20.37 [30] shows an example of a complete flow of the mask data after the IC design. The layout interchange format mentioned above, describes the layout in terms of geometrical elements such as rectangles and polygons.

In the 1970s a CAD system known as graphic design system II (GDSII) was developed by Calma Company and its format was called Calma Stream format [31]. This continues to be the industry standard today. An attractive feature of this format is that it uses "hierarchy" in its layout. This means that instead of describing every individual rectangle or polygon, these are described in form of cell and which are described in form of another cells, etc. This task is done by using a reference cell or even an array of cells as shown in Figure 20.38.

In recent years a new layout interchange format for ICs has been proposed by SEMI [32]. It is known as open artwork system interchange standard (OASIS). This is a step to overcome the shortcomings of GDSII by making the layout description file more compact. Moreover in the OASIS there is more room for future expansion of the format.

20.3.1.3.1 Mask Data Creation

Once the layers are derived then the appropriate corrections to the layout are applied to the data. For example OPC and PSM corrections are applied at this time. For example a simple line feature with just a few vertices may end up with segmented lines accompanied by lots of vertices as shown in Figure 20.39. This can cause significant increase in the file size and the run time. In some cases run time may even exceed the writing time.

After OPC is applied, data needs to be translated into the machine language which may differ from machine to machine. Moreover the data manipulation also involves operations such as: (a) Scaling, (b) Sizing of data, (c) Rotation, (d) Mirror Pattern, and (e) Tone Reversal as shown in Figure 20.40. At this time the data may also be corrected for electron-beam proximity effects or other such systematic errors that may be encountered during the processing of mask. During data preparation it is also necessary

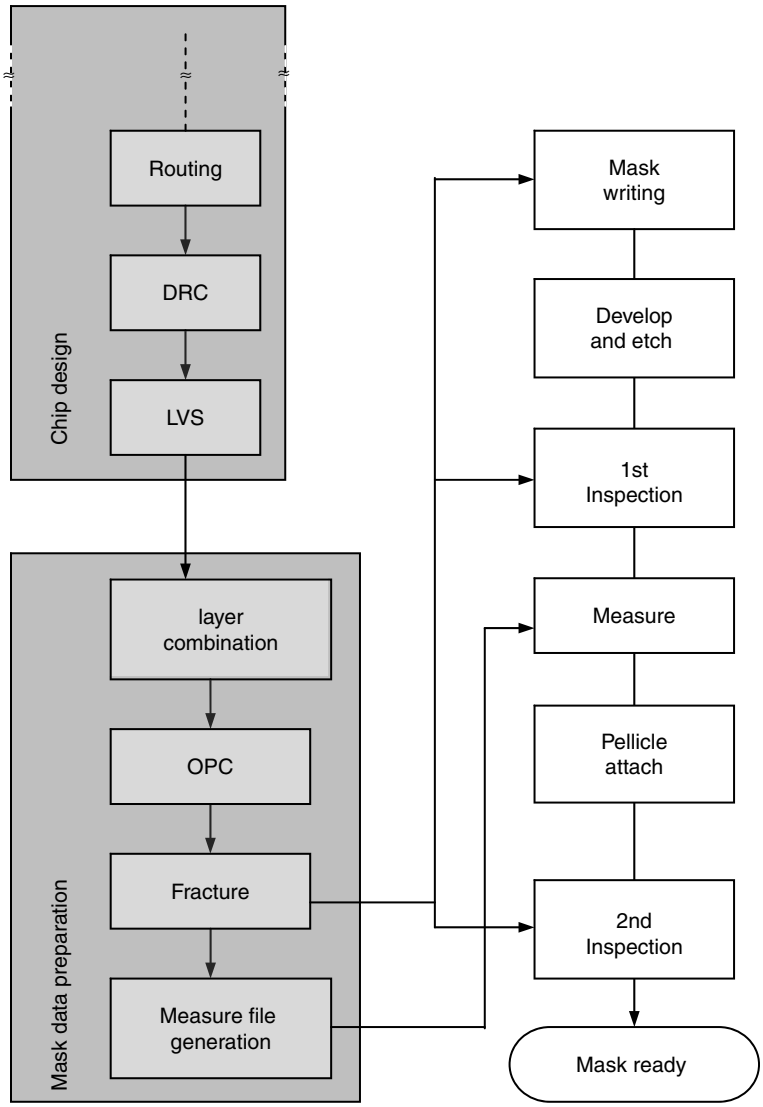


FIGURE 20.37 Data preparation and mask fabrication flow. (From van Adrichem, P. M. J., and C. K., Kalus, *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

to cerate a measurement set up file since there can be over 100 locations where CD and IP measurement are to be taken.

20.3.1.3.2 Jobdeck Creation

In mask writing it is quite common that the same pattern may have to be repeated at a number of places which can add to the data volume. It therefore would be more economical to create some sort of hierarchy, which is in fact the main idea of a jobdeck as shown in Figure 20.41 [30]. A jobdeck essentially is a table with references to pattern files and information on their placement on the layout. Besides this task a jobdeck can also contain various parameters and instructions for the mask writing tool, such as mirroring and scaling of pattern data, correction factors, beam setup parameters, sub-field or scan field placement information, etc.

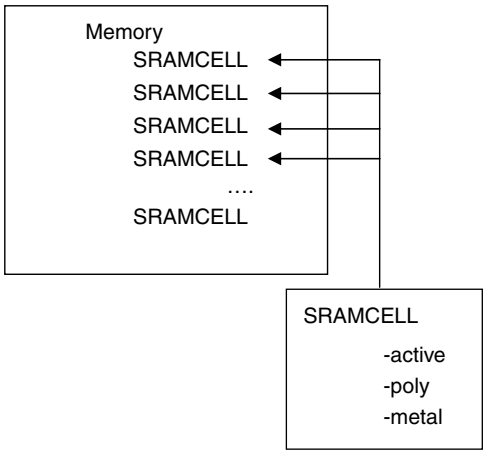


FIGURE 20.38 Graphic design system II format showing cell definitions and placements. (From van Adrichem, P. M. J., and C. K. Kalus, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

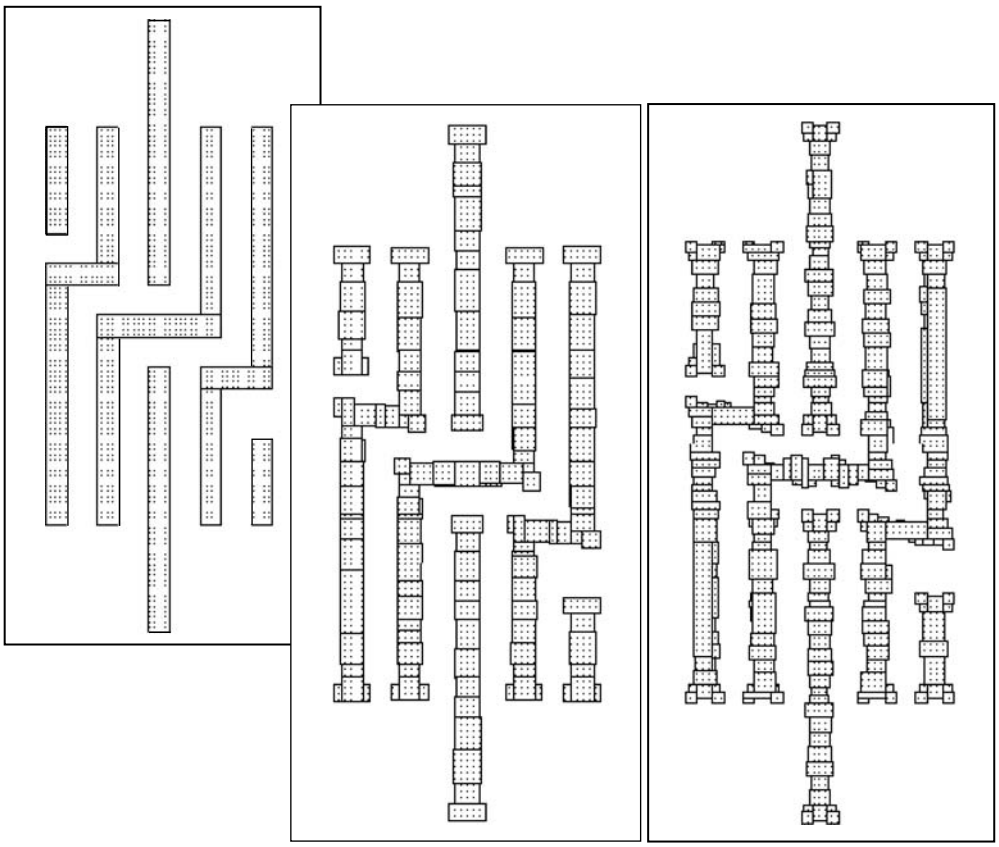


FIGURE 20.39 Optical proximity correction and aggressive OPC increase the number of vertices. (From van Adrichem, P. M. J., and C. K. Kalus, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

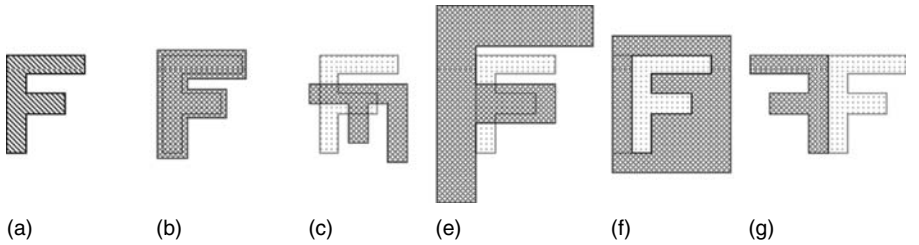


FIGURE 20.40 Common operation on data: like sizing, rotating, mirroring etc. (From van Adrichem, P. M. J., and C. K. Kalus, *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

20.3.1.3.3 Data Relationship with Machine Type

It was mentioned earlier there are various kinds of mask writing tools where they can differ in their formats, exposure system, and writing schemes, etc. Hence the data preparation needs to be tailored for the specific tool.

For example, in the case of raster scan the data is divided into rectangular blocks known as scan fields where the scanning takes place in one field at a time and it is inside these fields that a pattern or segment of a pattern are written by the scanning process. The size of the scanfields can vary from a few microns to several thousands of microns. Most formats use these scanfields to be able to cover the whole exposure area with a combination of stage travel and beam deflection. The stage moves the mask and places the right scanfield in the exposure area and the beam deflection "writes" the content of the scanfield. It would look desirable that an entire pattern or at least an isolated segment of a pattern could fit completely within one scan filed but in reality this may not be feasible. The way these scanfields are overlaid with the actual data can differ from machine to machine. In some systems the scanfield positions can be controlled, whereas in others these are hardware determined. It is possible that a pattern falls right on the interface of scanfields. In that case this pattern is written as two individual parts separated into two different scanfields and thus separated in time. This implies that the pattern fidelity such as line-width accuracy (CD-control) is a strong function of this scanfield stitching accuracy. Figure 20.42 shows an example of data that is divided into scan-fields.

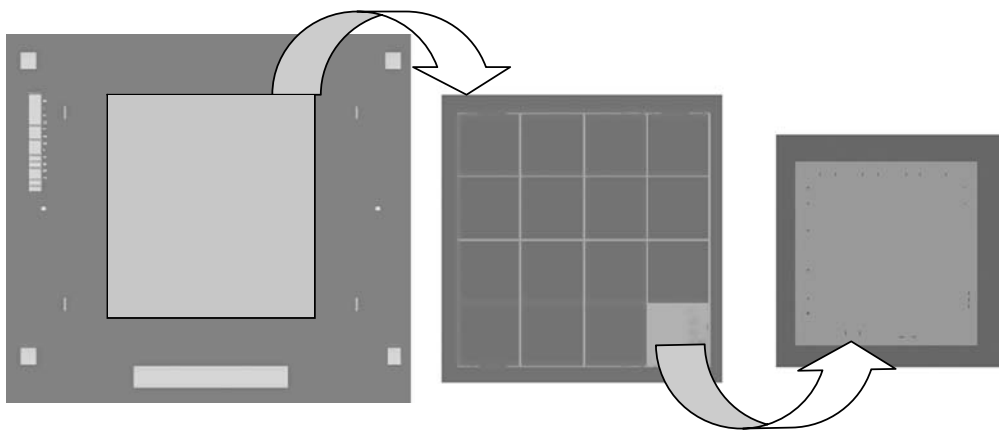


FIGURE 20.41 Using hierarchy when same patterns appear repeatedly. (From van Adrichem, P. M. J., and C. K. Kalus, *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

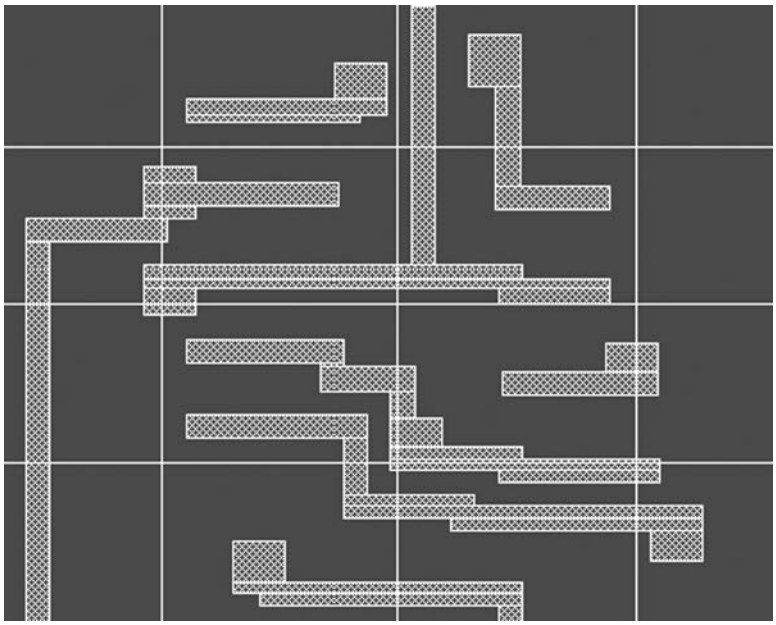


FIGURE 20.42 Same pattern being shared by different scan fields. (From van Adrichem, P. M. J., and C. K. Kalus, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

In case of vector scan the areas to be written are still partitioned into scanfields, but unlike in the previous case, here only the actual data-blocks, consisting of rectangles and trapezoids are exposed. The data can be broken into pixels before writing and can affect the write time in raster scan. The smaller the pixel size the more time it would take to write the mask. In vector scan, the write time is greatly affected by the area that has to be written, although here also, pixel size has the same role on the writing time as in the case of raster scan. Even then as compared with a raster scan tool, vector scanning tools do not usually show much of a higher writing time increase as a function of smaller pixel size.

In case of VSB it is able to expose a complete data element in a single shot. These data element can be rectangles and triangles or trapezoids.

Since the whole data pattern is divided into these elementary blocks for tools using this principle it does make a difference on how data is organized and sliced.

Following are some of the issues that are to be taken into considerations during MDP:

Pattern Complexities. Considering the continued advancement in technology like OPC and PSM these all are adding more demands on mask making in terms of in pattern complexity, higher data volume, and manufacturability. Figure 20.43 shows an example of OPC operation. The number of corner is increased that makes mask fabrication as well as inspection more demanding.

Grid Snapping. Grid snapping essentially occurs as a result of rounding off the data. The layout consists of polygons and path definitions. In case of polygon when all its vertices are on a certain grid, grid snapping will not occur. A *path* definition consists of a set of centerline coordinates plus a certain path width. Avoiding grid snapping in such a structure, not only requires the centerline coordinates to be on-grid but the width of the path needs to be a multiple of the double-grid as well (or the half-width a multiple of the grid). Under these conditions the perimeter of such a structure *can* be on this particular grid. However if there is a non-orthogonal section in a path, this perimeter is almost *never* exactly on this grid in the direct vicinity of this oblique section Figure 20.44. In other words in case of a non-orthogonal path definition in GDS, grid snapping is inevitable near the non-orthogonal edges.

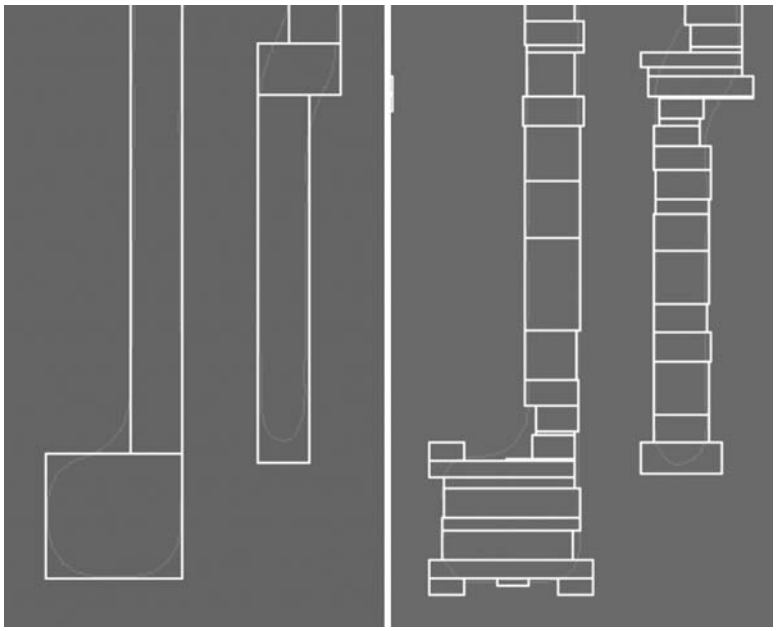


FIGURE 20.43 Optical proximity correction adding many jogs. (From van Adrichem, P. M. J., and C. K. Kalus, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

Data Slicing. There are several ways the input data can be sliced and difference in slicing can lead to differences in actual line-width accuracy. If a rectangle is divided into two elements, the total line-width error can be different from the case where the data is not split up Figure 20.45.

Scanfield Stitching. As mentioned earlier that it may be possible that some patterns may have to be shared by two scanfields. This may be a case when a data structure is placed exactly on a scanfield boundary and this structure could be written in two separate steps as shown in Figure 20.46.

Proximity Correction. Proximity effect when it was first mentioned in the chapter was in conjunction with the optical proximity effect that shows on wafer as an aberration from light diffraction. It is called

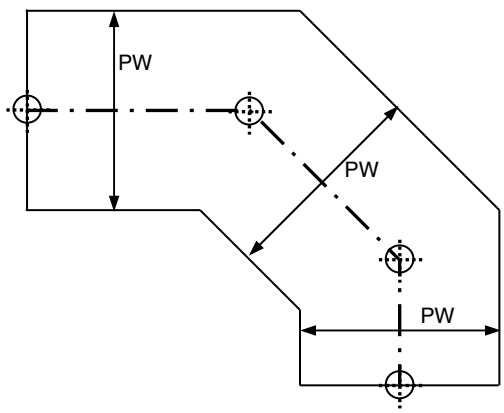


FIGURE 20.44 Defining path and grid snapping. (From van Adrichem, P. M. J., and C. K. Kalus, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

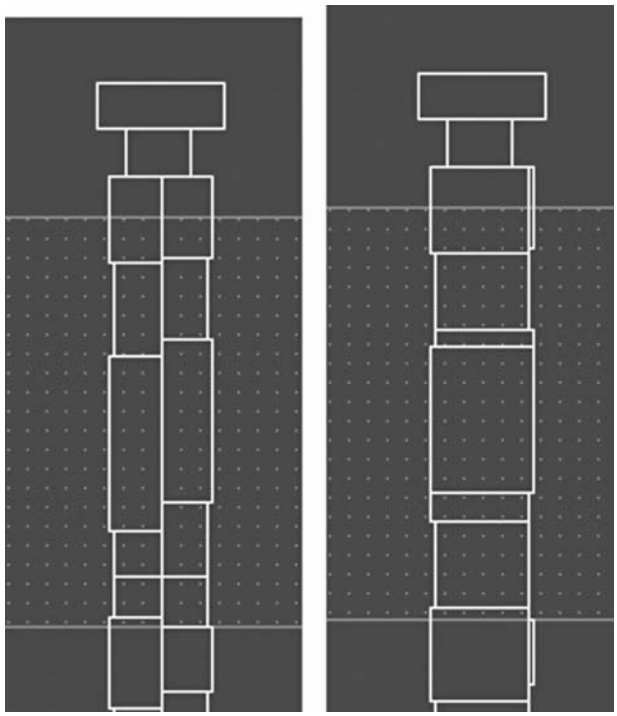


FIGURE 20.45 Slicing of data and line width error. (From van Adrichem, P. M. J., and C. K. Kalus, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

proximity effect because light diffraction from one feature affected the image of another feature in close proximity of the first feature. To counteract for this effect certain modifications in mask pattern are made and known as OPC.

Later the term "proximity" emerged again that is related to scattering of electrons in the resist during the e-beam writing on mask. Here again the scattering affected the features in the close proximity of the first feature. To compensate for these errors there have been methods like dose corrections, beam-shape corrections, and others.

A less computational way of applying proximity correction has been to compensate the back-scattered electrons with a background dose, which is written in as a second pass. This technique is

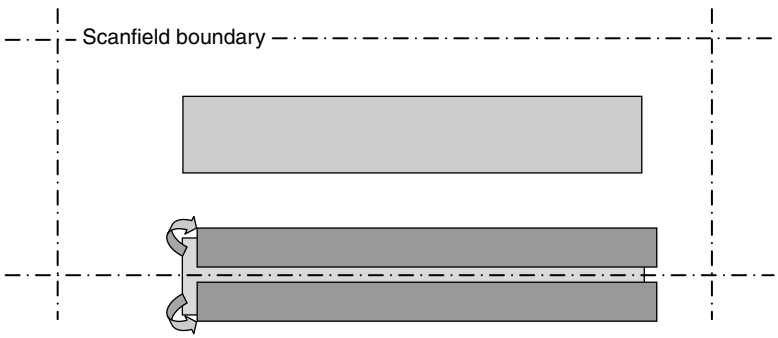


FIGURE 20.46 Scan field running through a data structure. (From van Adrichem, P. M. J., and C. K. Kalus, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

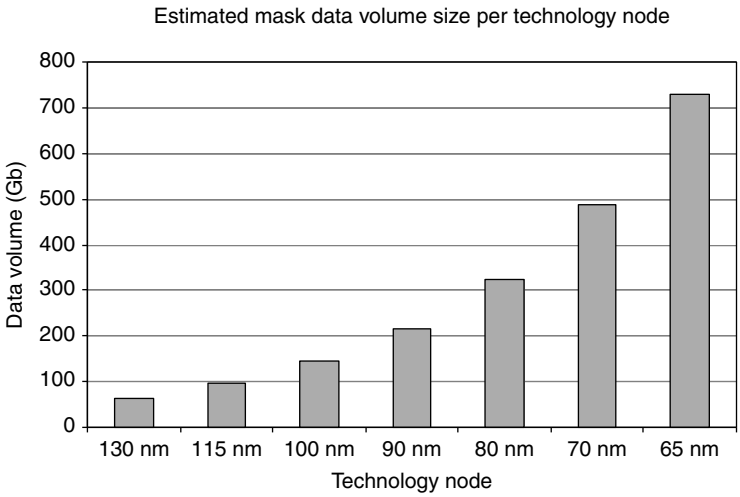


FIGURE 20.47 Mask data volume increase ITRS. (From van Adrichem, P. M. J., and C. K. Kalus, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 19–42. Boca Raton, FL, 2005.)

known as "GHOST" writing [33]. It uses a second writing pass with the inverted write data, which is written with a defocused beam. This wide beam is configured in such a way that together with the inverted image it yields a dose, which is the opposite of the back-scattered dose of the first writing pass. So the GHOST technique compensates the effect of back-scattered electrons by adding a background dose proportional to the inverse of the pattern density. The downside of this approach is increased cost in terms of time and resources.

20.3.1.3.4 Mask Data Processing Runtimes

With continued introduction of newer generation of products there has been a steep increase in MDP complexity. There are many reasons for such increase that include advancement in RET, more and more complex layer combinations (Booleans), increase of number of masks per process, use of different mask writing tools, and many more. All these individual items give rise to a data volume increase, and a huge increase in MDP run-time as shown in Figure 20.47. The total conversion time from DRC to read-to-write data experienced today is sometimes exceeding the total times needed to process the mask. On the other hand, the requirements for the total turn times for the final mask, so including mask making and data processing are getting tighter, which leave less time for the MDP.

There are different ways to reduce the MDP run-times, apart from the most obvious way, which is the use of a faster computer. Runtime improvement can be achieved by optimizing the flow of data manipulation. Some operations are inherently faster than others. When large intermediate files are generated and re-input to combine with other large files with some Boolean operation, such a Boolean operation can be quite costly in terms of CPU times [30].

Another way is by process parallelization. In this case one can have multiple CPU in a single machine. Another way is to split up the task to several machines and cluster. Even a combination of these approaches can be possible [34,35].

20.4 Materials and Processing

In the earlier sections we described the various types of photomask structures and machines used for their writing. In the following we describe materials and processes involved in mask making.

With the evolution in microlithography and mask design the mask substrate also has undergone many transitions.

In the early days of the semiconductor industry the mask substrate (also called mask blanks) used to be 2×2 in. sqr. The mask size has since then been growing, and at present the standard size is 6×6 in.[2] The thickness of the substrate has grown from 0.060 in. to the present 0.25 in. Today a 0.25 in. thick 6×6 in. substrate is referred as 6025 plates. The driving force behind this increase in the substrate size has been, among other factors, the large chip size and advancement in the step/scan exposure systems.

The starting material for substrate is glass plate sputtered with chrome based film that is then coated with photoresist. In this chapter the term mask blank or mask substrate means that glass plate has already been coated with the required material, e.g., chrome, MoSi, Photoresist, etc.

In the following sections a brief description of mask substrate and their processing is being presented.

20.4.1 Glass

The size of the glass plate is not the only factor of importance when considering substrate's structure. The glass material itself has undergone many changes.

In earlier days the standard material for mask substrate was used to be Soda Lime which was then replaced by a superior quality material known as White Crown for its reduced defects. Later on, when the thermal expansion of glass during exposure became an issue the White Crown was replaced with Boro-Silicate Glass which had a lower coefficient of thermal expansion (CTE). The next improvement was the introduction of Fused Silica (also known as Quartz) since the CTE of Fused Silica was even lower than that of Boro-Silicate. Besides its low CTE this Fused Silica material also exhibited better transmission at 365 nm referred as UV wavelength that was the standard in those days. This transparency of the material became more important when the industry moved from 365 to 248 nm and now to 193 nm illuminations. For the upcoming shorter wavelengths of 157 nm a F2-doped Fused Silica material with 76% transmission has been introduced [36,37].

Fused Silica has been the material for the leading edge technology for quite some time but with 193 nm exposure over an extended period the material has been found to exhibit color centers and compaction. The color center formation causes a low level of fluorescence at about 400 nm and compaction causes a small change in refractive index. However, these changes may affect the optics of the exposure systems but have no detrimental effect on photomask because the energy involved here is very small [38]. Schott-Lithotec a supplier of mask blank has also introduced a promising DUV Blank known as ZERODUR® with a spec of zero thermal expansion to meet the current requirements [39].

20.4.2 Chrome

At present the absorber on the glass is a compound of chrome consisting of Cr, N_2, O_2, and possibly other elements. The composition of the film varies from the bottom to the top serving different purposes. The bottom layer acts like a glue to improve the adherence of chrome to glass. The top surface acts as an anti reflective coating to minimize the undesirable reflection that may take place inside the system. The top and bottom layer of the film constitutes a very small portion of bulk of the chrome material that acts as the opaque film. A typical thickness of the film is 100 nm with an optical density of 3.0 which amounts to 0.1% (or less) transmission. As regards to amplified resists (AR), a coating based on a three layer Fabry–Perot structure with a reflectance of $<1\%$ has also been reported [40]. Smaller chrome thickness (59–73 nm) for improved performance has been explored and results are promising.

20.4.3 Molybdenum Silicide

Molybdenum silicide ($MoSiO_xN_y$) commonly known as MoSi films were first used to improve the adhesion to Fused Silica. Now MoSi is the key player in the structure of HT.PSM. The MoSi film is sandwiched between the glass and the chrome film. The MoSi has a tendency to flake and redeposit on

the mask during the processing of the substrate and can cause some yield loss. There are also issues with the exposure durability and the chemical durability of the cleaning process of the film.

Besides MoSi, there are other promising candidates like TaN/Si_xN_y, TiN/Si_xN_y, and $CrAlO_xN_y$ that are being looked at [40].

20.4.4 Other Substrate Related Topics

The homogeneity of glass in terms of its optical properties, e.g., refractive index, transmission, and birefringence has its direct impact on the CD uniformity [40,41]. A material developed by Corning quotes a low birefringence of <1 nm/cm and refractive index homogeneity as less than 4 ppm. The inhomogeneity of chrome and other phase related film on the glass can equally affect the CD uniformity.

Today's plates are considerably thicker compared to the earlier ones but because of their increased size the phenomena of "sag" and "distortion" when mounted on the exposure systems may still occur. Strains suffered by the plates under this condition can directly contribute to the IP errors. A deflection of 0.62 μm can give 40 nm IP error [38]. Image placement is also affected by the CTE of the glass. Even in Fused Silica a change of 0.08°C can change the IP by 10% of the allowed tolerance [38]. Plate flatness needs to be less than 1.0 μm. The spec on some of the plates has been quoted as low as 0.5 μm. Out of spec the flatness may affect CD uniformity and also affect IPs. Coated film may cause some stress on the glass and after the pattern is made some of the stress is released that can bend the glass causing an IP error.

20.4.5 Photoresist

Photomasks are written with electron beam as well as with laser beam and so the resists to be used for writing may differ accordingly. This section summarizes the different kinds of resist used for making photomasks [42].

For high resolution photomask for 90 and 65 nm nodes e-beam at 50 kV is used [43,44]. Although higher accelerating voltage can give better resolutions but it can also cause in resist heating resulting in reduced sensitivity [45–47]. Resists used for mask can be classified under the categories of non-CAR (*n*-CAR) and CAR where each type is further classified under the positive and negative types. These classifications along with their reactions are summarized in the following listing.

n-CAR Positive (chain scission)
n-CAR Positive (dissolution inhibition)
n-CAR Negative (cross-linking chain scission)
CAR Positive
CAR Negative

20.4.5.1 On-Chemically Amplified Resists (*n*-CAR)

n-CAR have been used since the early days of photomask lithography applications. The *n*-CARs are described on the basis of their functions as follows: positive resists that undergo chain scission, positive resists that convert dissolution inhibitors into soluble species, and negative resists that are based on crosslinking.

20.4.5.1.1 n-CAR Positive-Based on Chain Scission

These resists consists of polymer chains that undergo scission when irradiated with e-beam resulting in the reduction of their molecular weight. This increases the solubility of the polymer in an organic solvent. The difference in the solubility between the exposed and un-exposed part gives rise to the formation of pattern after they are treated with the solvent. Most commonly used polymers for mask writing have been PMMA, and poly butane sulfone (PBS). In terms of their sensitivities, PBS meets the required sensitivity of 1 $\mu C/cm^2$ at 10 kV which is 10 times higher than the sensitivity of PMMA. Both resists have been used in the industry for decades but neither has exhibited adequate resistance to dry etch [42,48].

Another resist ZEP 7000 developed by Dai Nippon has been widely used by the industry. Its sensitivity is about 8 $\mu C/cm^2$ at 10 kV and it provides adequate resistance to dry etch [49].

20.4.5.1.2 n-CAR Positive-Based on Dissolution Inhibition

There is another kind of positive tone resist that is based on dissolution inhibition. Two-component diazonaphtoquinone (DNQ)-novolak were developed for 435–365 nm lithography for wafers.

20.4.5.1.3 n-CAR Negative-Based on Cross-Linking

There are negative tone *n*-CARs that are based on cross linking. These are known as crystal-originated pits (COP) based on the copolymer of glycidyl methacrylate and ethyl acrylate [49]. COP however has the drawback of swelling and poor etch resistance. Another negative tone resist is hydrogen silsesquioxane (HSQ). In this case heat and e-beam exposure [50,51] cleave the SiH bonds enabling the formation of a SiO cross-link network which is insoluble in standard tetramethyl ammonium hydroxide (TMAH) or KOH developers. Hydrogen silsesquioxane has been used to print 30 nm structures for imprint lithography masks [52], and 10–30 nm structures for direct write silicon applications [53,54].

20.4.5.2 Chemically Amplified Resists (CAR)

Chemically amplified resists, give good resolution and also show very high sensitivity to the incoming exposure energy. These types of resists are based on radiation-induced generation of a catalytic species, usually strong acid that brings about multiple chemical transformations that change the solubility of the polymeric matrix of the resist in developer (e.g., aqueous TMAH solution). Chemically amplified resists are available in positive as well as in negative tones and these can be used for optical wavelength (257, 248, 193, and 157 nm), and also for e-beam, EUV, and x-rays [42].

20.4.5.2.1 CAR Positive Tone

Currently most commercially available CARs are positive tone resists. First CARs were based on protected poly (*p*-hydroxystyrene) and designed for 248 nm lithography [55–57] and later developments were extended to cover for shorter wavelengths of 193 and 157 nm as well. For 257 nm laser applications, the resist for 248 nm was found to be quite adequate. Interestingly enough, these resists also turned out to be quite suitable for e-beam exposure as well. There are some CARs specifically designed for e-beam applications such as FEP-materials developed by Fuji-Arch, REAP series from TOK, and the DX family of resists from Clariant.

20.4.5.2.2 CAR Negative Tone

Most negative tone CARs are based upon cross-linking like their *n*-CAR counterparts are. The negative tone CARs offer high sensitivity similar to the positive tone CARs. Here the unexposed polymer film is soluble in aqueous base, and the acid catalyzed reaction enables covalent bond formation between polymer chains and multifunctional cross-link agents. The resist was originally developed for 248 nm lithography and later extended to e-beam and 257 nm laser application for mask making.

20.4.6 Processing

The processing of mask involves steps starting from resist development down to the final cleanup until the pellicles are mounted on it. For the conventional COG structure the mask processing involves the following basic steps.

1. Developing the mask after exposure to laser or e-beam.
2. Pattern Transfer: Etching of chrome uncovered after resist is developed.
3. Stripping the resist
4. Final cleanup

Departure from the above process can occur for masks with more complex structures. Depending on the type of resists some bake operations may be required during the processing.

Typically, the COG mask blanks are coated with chrome and resist film at the supplier's site before they are delivered to the mask shop. At the mask shop the blanks are exposed on laser or e-beam writers and then run through the appropriate developer resulting in the appearance of resist patterns on the plate.

20.4.6.1 Resist Coat and Develop

As mentioned earlier resists can be positive tone or negative tone and which determines what part of the pattern will be developed out and which will not. The resulting resist pattern then is transferred onto the underlying layer which will be chrome in the current example.

Although most blanks coming into the mask shop are already coated by the supplier, there are cases where at some point in the process flow, further coating of resists may have to be done at the mask facility. The procedure is called "spin-coat" where the resist is delivered onto the surface of mask mounted on a chuck that is then spun at a speed of a few thousands rpm. The chuck when spun at a prescribed speed causes the resist to spread with the formation of a uniform film thickness across the mask surface. A typical spec on resist thickness is 300–450 nm with a uniformity of ± 3 nm. The coated plate is then baked in order to remove solvent. After this the plate is ready for exposure. Following the exposure, the plate is developed.

There are two techniques commonly employed for developing masks after exposure.

One technique is the immersion method where the mask is dipped into a tank of developer where the chemical reaction takes place. An advantage of this method is that it can readily be adapted to batch processing where several plates can simultaneously be processed. The method however has a drawback of contaminating the mask that later may require an additional clean ups. The other, more common practice is "spin-develop" with mechanism similar to "spin-coat" mentioned earlier. Here the plate always sees fresh developer at the beginning of each cycle and hence is less prone to contamination.

The uniformity of development across the plate can have an effect on CD uniformity and hence due to the nature of spin-develop a degree of radial CD variation on the plate is possible.

20.4.6.2 Pattern Transfer: Etching of Chrome

After the resist has been developed, the next step is to transfer the resist pattern onto the underlying chrome film that is at this stage is unprotected after the removal of resist. This task is carried out by the etching of the unprotected chrome. However, at this point it is important to examine the plate to make sure the resist is completely developed out. With certain types of resists an iterative process of CD-measurement and re-development may be necessary until the required CD is achieved. In some cases traces of resist known as "scum" are left behind in the open windows that can interfere with the etching of chrome. These scums can be removed by a quick exposure to oxygen plasmas. The process is called "de-scum." The next step is then the etching of chrome, which historically has been done with wet chemicals that are regarded as a part of wet processing. In recent years, due to increased demands on the tolerance of shrinking features, many plasma etch processes known as dry etch processes are also being employed.

20.4.6.2.1 Wet Etching of Chrome

The processing stations for wet etch can be similar to that of develop stations namely immersion tanks or spin stations. The chemicals commonly used for chrome etch are Cerric Ammonium Nitrate and certain acids that include Perchloric, Acetic, Nitric, and Hydrochloric acids [36]. Due to the liquid nature of the chemicals the wet process tends to be isotropic and causes certain degrees of under-cuts. These under-cuts however turn out to be helpful in minimizing the effect of slope that the resist profile exhibits towards its edges.

20.4.6.2.2 Dry Etching of Chrome (or Other Underlying Material)

As the feature sizes are getting smaller and the tolerance of their size are getting more demanding, the industry is moving toward dry etch process. Dry etch process which is more or less an anisotropic and hence it requires very little or no process bias. Dry etch can also meet the stringent tolerance on CDs as required by the state of the art designs.

Considering a 100 nm technology node it would seem that the mask feature would be 400 nm. However when OPCs are involved the 4:1 rules breakdown and in such cases mask feature need to be significantly below 200 nm. At present 100 nm or even smaller features (on mask) are being pursued. Dry etching involves use of plasma (a mixture of electrons, ions, and various neutral species). In today's vocabulary dry etch has become synonymous with plasma etch.

20.4.6.2.2.1 Plasma Reactors

There are various types of reactors for the creation and application of plasma that are used for the etching of chrome or whatever underlying material there is to etch. Some such examples of the etching systems are given in the following:

Ion Milling. Here ions are accelerated toward the target where it is the mechanical impact of the ions rather than any sort of chemical reaction that does the etching of the chrome film.

Reactive Ion Etching (RIE) and Magnetic Enhanced RIE (MERIE). In this case a reactive species in the plasma chemically reacts with the target to increase etching rate. The composition of chrome etching is $CH_2Cl_2 + O_2$ [36].

Inductive Coupled Plasma (ICP). This is a low pressure and a high density plasma. It gives improved CD control and uniformity. Inductive coupled plasma is also good for low defect counts.

Plasma Applications and Processes. Although in the case of COG or binary masks it's only the chrome film that need to be etched, but with the emergence of phase shift masks, new processes are being fashioned that can etch materials such as MoSiON and others. There are also chromeless mask where features are etched into the quartz which is now being done with plasmas.

In one example of MERIE the etch parameter for Cr and MoSiON have been cited as the Cl_2/O_2 with Gas Assisted Etching (GAE), or for Cr, the composition involved Cl_2/O_2 with GAE and for MoSiON the composition was CF_4/O_2 for MoSiON. The GAE increases the etch selectivity 1.8 times higher than without GAE [58].

There are a number of factors that affect the plasma etching and need to be addressed.

An important factor is chrome loading, that is the amount of chrome on the mask also affects several plasma etch responses, e.g., resist selectivity, Cr etch rate, overall CD uniformity, and within mask uniformity [59]. During the dry etch of chrome a certain amount of resist is lost and appears as re-deposited polymers and debris on mask surface adding to increased defect counts. The resist lost can also affect the CD uniformity and etch bias. The phenomena of resist loss are related to poor selectivity. The objective is then to minimize this resist loss by improving the selectivity. The CD control in uniformity and etch bias show opposite trend lines with standard chemistry of $H_2/Cl_2/O_2$. Increasing the Oxygen flow can improve the uniformity but it also decreases the selectivity. In order to overcome the limitations of the two opposite trends, what is needed is to develop a process with an improved selectivity to photoresist and with reduced dependence on O_2 flow. There have been chemistries proposed that could provide this benefit. Hydrogen and carbon containing gases are considered as the promising alternatives. Several gases proposed are H_2, HCl, and NH_3 and carbon containing gases to promote selectivity are C_2F_6, CCl_4, C_3F_8, CHF_3, CH_4, and CF_4–H_2, etc., [60].

Another work, also on chrome etch, reports on achieving 90 nm features on masks using ICP reactor. To feature this small, it requires a number of process optimization other than just plasma. Such as the type of resist, processing, and the writing scheme, etc., [61]. Unaxis, another major supplier of plasma etch systems has introduced its Unaxis MASK ETCHER® IV that will address 100 and 65 nm technology node [62]. The uniformity of trench depth in quartz etched by this system was within the noise level of the atomic force microscopy (AFM) used for measuring the trench depth.

As mentioned earlier there are phase shift masks that require etching of the quartz to the right depth. Here, after the opening of chrome windows, it is the glass (quartz) that is to be etched. In the case of quartz etching there is no under-layer that can be used as an etch stop. In such cases the technique is to etch for a predetermined time that can be guaranteed for the desired depth. The work referred here utilized ICP source with gas composition as CHF_3:CF_4 [63].

20.4.6.2.3 Resist Stripping and Cleaning

Cleaning of photomask starts from the stripping operation where the unwanted resist after chrome etch is to be removed. However, simply stripping of the resist does not result in perfectly clean mask. There can be defects arising from number of sources where some may be as common as watermarks whereas others may be more subtle and extremely small in size. These particles and other contaminants can adhere to mask surface by Van der Waals or electrostatic forces and can be detected only by sophisticated techniques.

20.4.6.2.3.1 Wet and Dry Cleaning Processes

In general, the cleaning operations fall under the class of wet processes where masks are cleaned with some type of liquid solution; or it can fall under the dry process where the mask is exposed to a plasma environment while subjected to high energy photons for its cleaning. At present, most operations are carried out using wet processes although dry processes are beginning to emerge at many facilities.

20.4.6.2.3.2 Cleaning with Wet Processes

The wet process where the mask is subjected to some form of liquid treatment can be further classified as mechanical or purely chemical in its nature. Example of mechanical treatment is the scrubbing of masks with specially designed brush or sponge, whereas in the other case the mask is immersed into a tank of chemicals that cleans the mask. Then there are also techniques that employ high pressure spray cleanup that can be seen as to take the advantage of mechanical impact of the spray as well as reactions with the chemicals to dislodge the contamination from the mask surfaces.

20.4.6.2.3.3 Chemistry of Wet Process

Most of the wet processes involve a mixture of H_2SO_4 and H_2O_2 in the ratio of 4:1 used at 90°C commonly known as Piranha Clean and is primarily used for removing the resist and heavy organic material. It works as an oxidant and attacks the hydrocarbons [38]. Another material used for mask cleaning is a mixture of H_2O, H_2O_2, and NH_4OH in the ratios of (5:1:1) used at room temperature and known as RCA Standard Clean-1 or simply SC-1. This chemistry was designed for removing traces of organic impurities from the mask surface by [38] solvating action of NH_4OH and the oxidation capability of the H_2O_2. The NH_4OH also serves as a complexant for many metallic contaminants. In this case, the peroxide in the solution oxidizes the surface and then the ammonium hydroxide dissolves this oxide. Although this sequential growth and etching of the surface helps in the removal of particles and it also results in the undesired micro-roughening of substrate. Recent research has shown that lowering the NH_4OH concentration ratio to 0.01–0.25 greatly reduces the micro-roughening while retaining the particle removal efficiency of the SC-1 clean [38].

Newer techniques like ultrasonic and megasonic cleaning are also becoming quite prevalent.

20.4.6.2.3.4 Cleaning with Dry Processes

Dry cleaning is mainly associated with use of plasma that reacts with the contamination resulting in a by-product which is then flushed out by the flowing gas. There is also another area of dry processing known as laser-assisted-cleaning. Developed by Radiance Services Company [64,65] the process uses high energy photons that can break bonds that hold particles to surface without any surface damage. The system utilizes flowing gas that sweeps the particles away from the mask area. Since the process is dry and uses no water or toxic chemicals; benefits of the technology may include a reduced need for deionized water, chemical handlers, and waste treatment systems in semiconductor facilities. The use of UV radiation has also been known to help strip resist from mask by weakening the bonds of remaining particles after the first strip process.

20.4.6.2.3.5 Cleaning with Semi-Dry Processes

Besides the wet and dry cleaning processes there are processes that can be classified as semi-dry process. In one case, by employing vapor of some liquid (water, isopropanol, and ethanol), a layer of water is deposited on the mask after which the particle is "hit" with a laser beam. The heated water turns into

steam and lifts the particle off the mask surface and it is then carried away by a stream of gas flow across the mask surface. The liquid used could be water [66].

20.4.7 Pellicles

A pellicle is a thin and transparent membrane stretched over a sturdy frame which is then mounted on a mask to protect its surface from foreign particles and other such contamination. During the printing process, the image of any particle on the pellicle film will be out of focus on the wafer plane and will not print. The use of a pellicle in an optical projection system is illustrated in Figure 20.48 [67]. A pellicle must have a good light transmission as well as a long term transmission stability.

The transmission of a thin film is dependent on the film thickness, light wavelength, incident angle, and light absorption of the film. For a normal incident of light on a non-absorptive thin film, the maximum transmission of the film happens when a film thickness is an integer multiple of an optical half wavelength, i.e., the half wavelength of the light divided by the refractive index. That is when thickness t equals $k\lambda/2n$, where k is an integer, λ is the wavelength, and n is the refractive index. And the transmission is minimum when the film thickness is at the optical quarter wavelength from a transmission maximum. That is when $t = k\lambda/2n + \lambda/4n$. In other words, when the refractive index $n = 1.5$, the minimum transmission turns out to be 0.84.

Anti-reflective coating was introduced to improve the transmission and reduce the sensitivity of transmission to the film thickness variation. Later on multiple-layer anti-reflective coating was also introduced [68]. Initially Nitrocellulose was used for film material which worked well for g-line (436 nm) and i-line (365 nm) wafer steppers, but because of its increased absorption below 350 nm it was not suitable for DUV stepper/scanner systems. Cellulose esters, such as cellulose acetate and cellulose acetate butyrate, have shown good transmission above 300 nm while amorphous per-fluoropolymer materials,

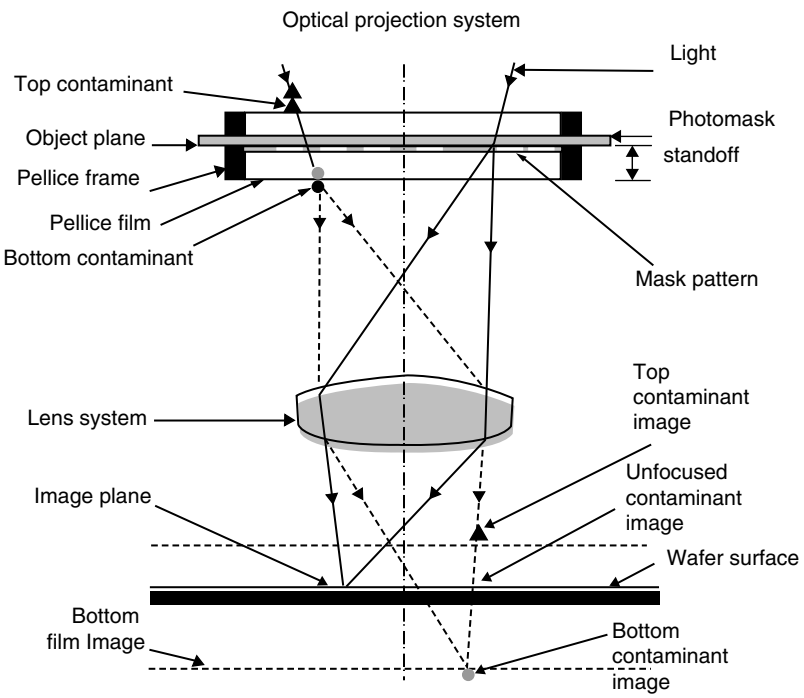


FIGURE 20.48 Use of a pellicle. (From Yen, Y- T., G. B. Wang, and R. Heuser, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 395–410. Boca Raton, FL, 2005.)

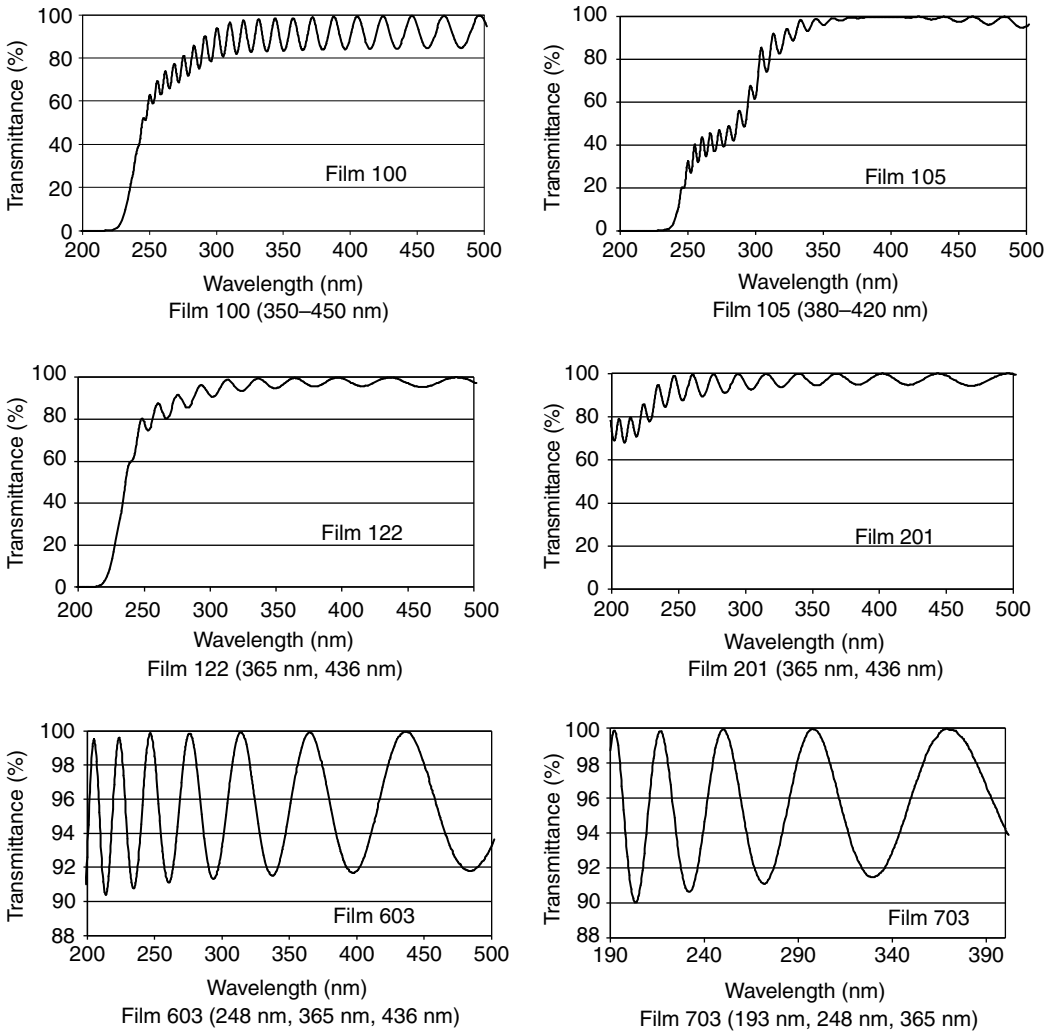


FIGURE 20.49 Typical transmission curves. (From Yen, Y- T., G. B. Wang, and R. Heuser, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 395–410. Boca Raton, FL, 2005.)

such as Teflon AF® (DuPont™) or Cytop® (Asahi Glass Co. Ltd) can be used for 248 nm and 193 nm steppers. An example of the different types of film and transmission curves is shown in Figure 20.49 [67]. A typical pellicle is shown in Figure 20.50 [67].

20.5 Photomask Qualification

20.5.1 Metrology

Photomask contains pattern of circuits that are then imaged on wafers. The width of certain features is very critical for adequate performance of the circuit. These features known as CDs or critical dimensions must be measured on mask and later on, on wafers. Besides the width of features it is equally important that all features are at the right locations since the circuits being made on wafer consists of many layers which

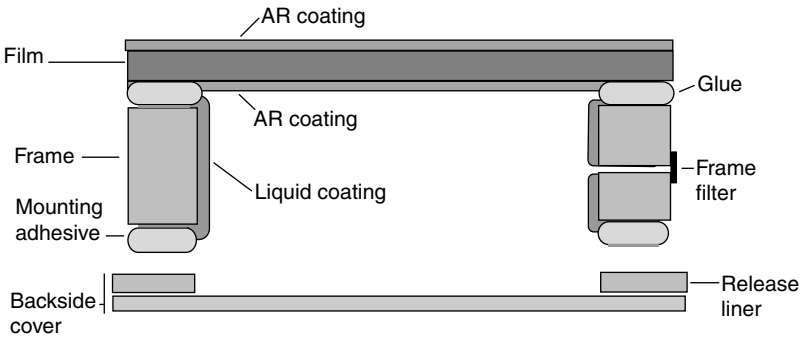


FIGURE 20.50 Cross section of a micro lithography Inc. (MLI) pellicle. (From Yen, Y- T., G. B. Wang, and R. Heuser, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 395–410. Boca Raton, FL, 2005.)

should match to each other. This can be possible if all features are at their right location as prescribed by their design rule. Measurement of the feature positions is known as image placement (IP) measurement or simply IP measurement. A machine to carry on such a task not only measures the coordinate of features but also compares those values with their design values and outputs the results accordingly.

Hence metrology in masks making can be classified under CD metrology and IP metrology as described in the following.

20.5.1.1 CD Metrology

20.5.1.1.1 Optical CD Metrology

Among the number of ways for CD measurement the optical imaging techniques have been the most commonly used since the very early days of semiconductor manufacturing. Optical imaging techniques used in the past have been addressed to individual features. However, in recent years, there has been growing interest in measuring a group of lines.

One such method is optical based diffraction technique. The technique does not directly measure any specific feature but is based upon rigorous coupled-wave analysis (RCWA) for periodic grating structures which are normally put along with other test patterns for the circuit. The technique is capable of measuring the CDs of grating structures down to approximately 40 nm. It does not require high vacuum or any kind of environmental chambers like some other technique do, and it can be readily integrated into the existing process tools. The technique however does require that the specimen to be measured be in form of a grating.

For the current technology the target structures on wafer are limited by illumination and should have their dimension not less than 50 μm × 50 μm with a nominal pitch of 180 nm or greater. On mask, these numbers however are magnified by 4×. This means that for 100 nm wafer level lines, the typical mask measurement would be 400 nm. It may appear that at these larger dimensions the traditional imaging methods may suffice, however the optical CD metrology (OCD) method can provide profile information that can greatly enhance process development work. For example, obtaining cross-section information by traditional imaging methods is a time consuming and is a destructive process, but here such information can be readily obtained without resorting to some destructive technique.

There are a number of OCD techniques such as Scatterometry, Spectroscopic Ellipsometry, Normal Incidence Spectroscopic Reflectance, and Normal Incidence Spectroscopic Ellipsometry. All these methods require a periodic structure (gratings) to diffract the oncoming beam. The analysis of the diffracted beam can then provide the information on CD features. The analysis requires the modeling of the diffracted light based upon the optical properties and the structure of the grating. These models

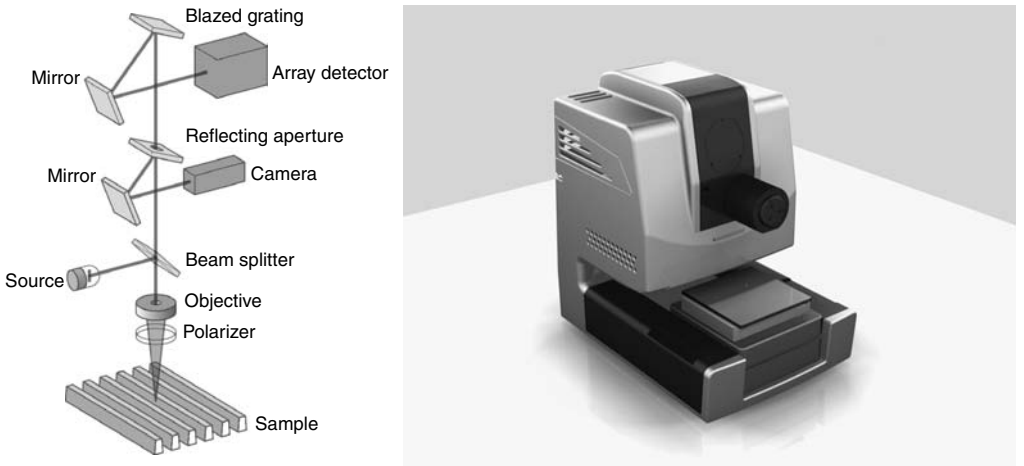


FIGURE 20.51 Normal incidence ellipsometry (NANOmetrics, Inc.). (From Hoobler, R. J., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 433–55. Boca Raton, FL, 2005.)

employ complex theories such as RCWA in conjunction with experimental data to help generate data and library for later uses.

Figure 20.51 [69] shows a set-up for Normal-Incidence Spectroscopic Ellipsometry along with associated NANOmetrics 9010M OCD system.

This is another technique where the polarization of the incident light is set with respect to the periodic grating structure on the sample allowing several different modes of data acquisition. For analysis of complex structures, phase information can be obtained by acquiring reflectance data at a polarization of 45° with respect to the grating structure.

The normal-incidence spectroscopic ellipsometer maintains much of the simplicity in mechanical design found in a standard reflectometer. Following are some examples that can be applied during the process development.

20.5.1.1.1.1 Resist Grating on Chrome

Figure 20.52 shows a well-defined grating structure (A) and a poorly defined grating structure where the photoresist material has not been totally cleared (B). A comparison of two spectra shows the distinct spectral changes due to the grating structures.

20.5.1.1.1.2 Chrome Grating Structures

Here a typical structure for chrome grating on quartz can be simulated to observe the sensitivity to changes in CD. Figure 20.53 shows the calculated reflectance spectra for a chrome grating on a quartz substrate.

20.5.1.1.2 SEM CD Metrology

Scanning electron microscope as the name itself implies was originally developed as a microscope. The usage of SEM technology is now extended to the measurement of feature width known as CDs.

The SEM provides higher resolutions than is possible by other optical techniques.

In 1984, Postek [70,71] presented some early work on SEM metrology of photomasks. At that time SEM metrology was at its early stages. Table 20.3 [72] summarizes later updates on the technology [73]. Detail descriptions of SEM Metrology have been addressed by Postek and Vladar [74], Postek and Larrabee [75], and Postek and Joy [76]. In a SEM, a finely focused beam of electrons is raster scanned over a small area of interest on a specimen.

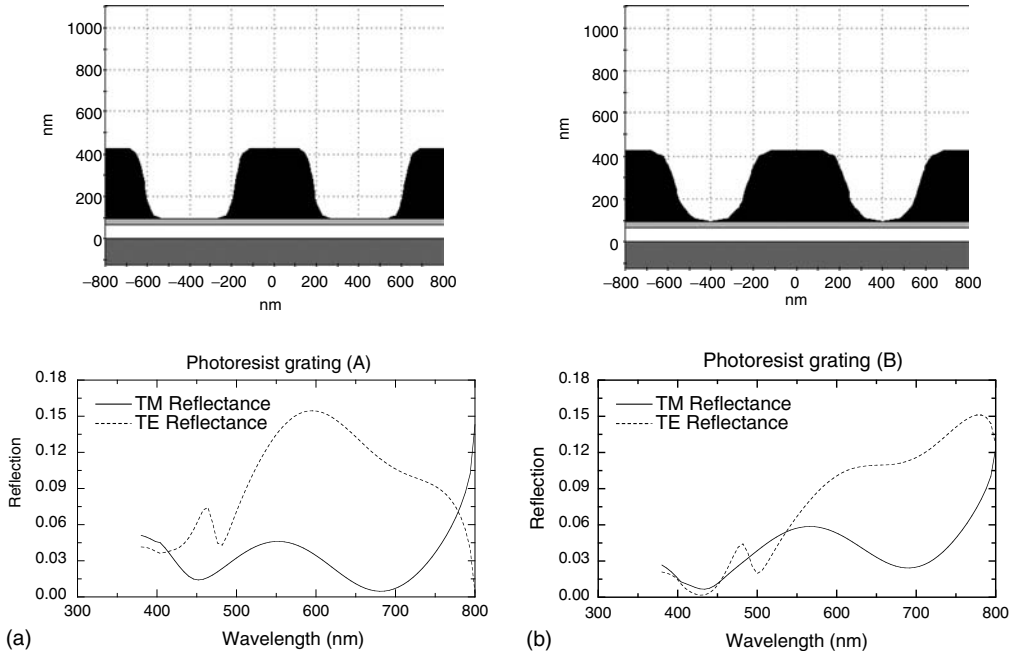


FIGURE 20.52 Spectra for well resolved (a) and poorly resolved (b) grating. (From Hoobler, R. J., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 433–55. Boca Raton, FL, 2005.)

The primary beam consists of electrons originating from its source and then accelerated toward the specimen. The accelerating voltage typically is between 0.1 and 30 kV. As the beam moved down the column its diameter undergoes demagnification from a few microns to a few nanometers. The choice of the accelerating potential depends upon the application and number of factors such as, resolution, nature of specimen, etc.

The majority of CD metrology of photomasks is currently done under "non-destructive" SEM conditions. Non-destructive inspection in a SEM implies that the specimen is not altered before insertion into the SEM. Historically, scanning electron microscopy was done at relatively high (typically 2–30 kV) accelerating voltages in order to obtain both the best signal-to-noise ratio and image resolution. At high

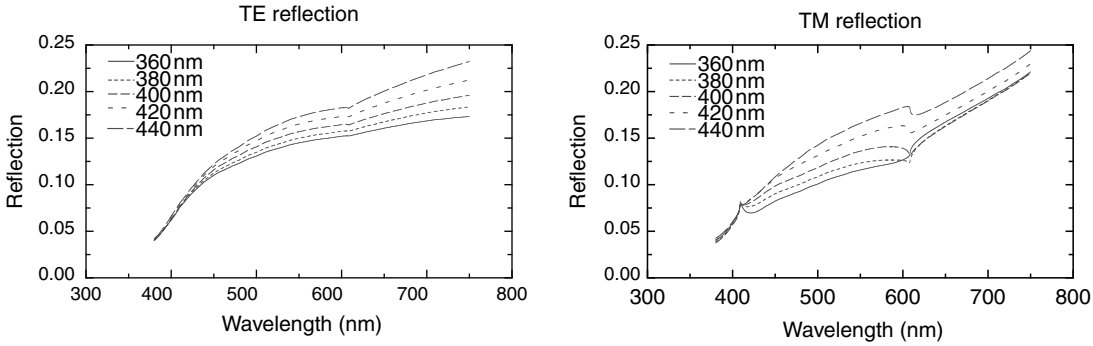


FIGURE 20.53 Spectra for chrome gratings show variations with respect to line width. (From Hoobler, R. J., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 433–55. Boca Raton, FL, 2005.)

TABLE 20.3 Comparison of Selected Characteristics of Scanning Electron Microscopy (SEM) Photomask Metrology Instruments—Past and Present

Characteristic	1985 Instrument	Current Instrument
Instrumentation	Modified laboratory SEM	Modified Dedicated CD-SEM
Lens Technology	Flat, 45° or 60° pinhole final lens	Extended field and other
Automation	Non-existent	Common
Electronics	Analog/digital	Fully digital
Image averaging	Rudimentary	Sophisticated and built-in
Scan rate	Slow scan	TV rate
Electron source	Lanthanum hexaboride	Field emission
Linewidth measurement	Rudimentary	Model and library-based
Throughput	Poor (manual)	Vastly improved
Charging	Problem	Acceptable solutions
Sample size	Broken mask/full mask with difficulty	Full mask
Sample contamination	Marginal	Improved
Signal-to-noise ratio (S/N)	Poor	Greatly improved
Cost	$150K	$2.5M+

Source: From Postek, M.T., *Handbook of Photomask Manufacturing Technology*, ed. Rizvi, S., Taylor & Francis Group, 457–97. Boca Raton, FL, 2005.

accelerating voltages, non-conducting samples require a coating of gold or a similar material to avoid charging effect that interferes with the incoming beams. Unlike the earlier systems today's machine can accommodate the entire mask and the machines are run at lower voltage to avoid charging and thus they do not need to be coated with gold or any such material.

Low acceleration voltage operation for production and fabrication of photomasks remains of great interest to the semiconductor industry [77–80]. At low accelerating voltages it is possible to inspect photomasks and in-process wafers in a non-destructive manner. Low accelerating voltage operation is generally defined as work below 2.5 keV—generally within a range of about 0.2–1.2 keV.

The interaction of an energetic electron beam with a solid results in a variety of potential "signals" generated from a finite interaction region of the sample [81]. The most commonly used signals of SEM are the secondary electron (SE) and backscattered electron (BSE) signals. The distribution and general intensity of these two signal types is shown in Figure 20.54. The size of the interaction region is directly

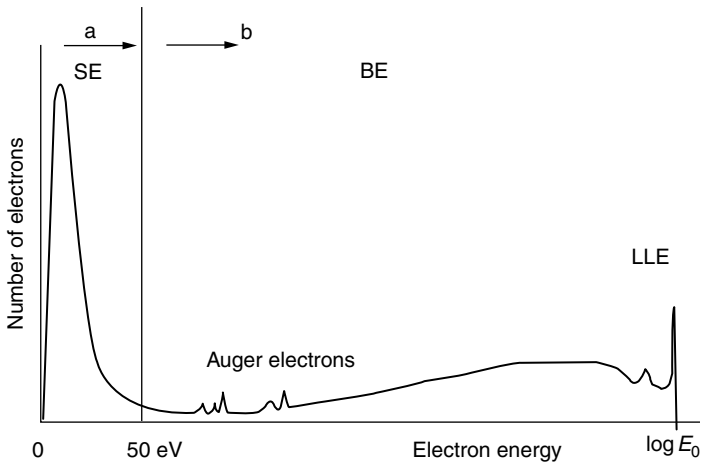


FIGURE 20.54 Distribution and intensity of some of the typical scanning electron microscopy signal types. (From Postek, M. T., *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 457–97. Boca Raton, FL, 2005.)

related to the accelerating voltage of the primary electron beam, the sample composition, and the sample geometry. The signals produced within the interaction region leave the sample surface which can then be used for imaging [72].

The appearance of a scanning electron micrograph is such that its interpretation seems simple. However, it is clear that the interaction of electrons with a solid is an extremely complex subject. Each electron may scatter several thousand times before escaping or losing its energy, and a billion or more electrons per second may hit the sample. Statistical techniques are appropriate means for attempting to mathematically model the interactions. The most adaptable tool, now, is the so-called Monte Carlo simulation technique. In this technique, the interactions are modeled and the trajectories of individual electrons are tracked through the sample and substrate. With the potential of much less than 100 nm linewidths and high aspect ratio structures, the SEM remains an important tool for CD metrology. An accurate metrology with this instrument requires the development and availability of traceable standards.

20.5.1.1.3 AFM CD Metrology

Today's high end masks especially Alt.PSM have patterns that are composed not only of two-dimensional features but also of trenches and dual trenches in quartz that need to be accurately measured. Moreover, at nano scale it becomes important to have precise information of three-dimensional profile of features. In some Alt.PSM there are over hangs that are specially designed and they need to be measured in order to make sure they meet the required specs. Such measurements can best be made using AFM techniques.

AFM can be seen as a refined version of early days machines like Dek-Tak or Tally-Step used for measuring profile and surface roughness of very small features. Figure 20.55 [82] shows a schematic of the workings of an AFM. The tip of the AFM almost touches the surface but not quite. Depending on their design and structure there can be various types of forces of interaction between the tip and the artifact being probed. Forces of interaction as shown in Figure 20.56 [83] range from micro to pico Newton.

All scanning probe microscope (SPMs) consist of a probe tip, a sensor that accurately locates the vertical position of the tip, a feedback system that controls the vertical position of the tip and a piezoelectric scanner that moves the tip relative to the sample in a raster pattern. A computer system drives the scanner, measures the data and converts it into an image. The tip is linked to a cantilever. The

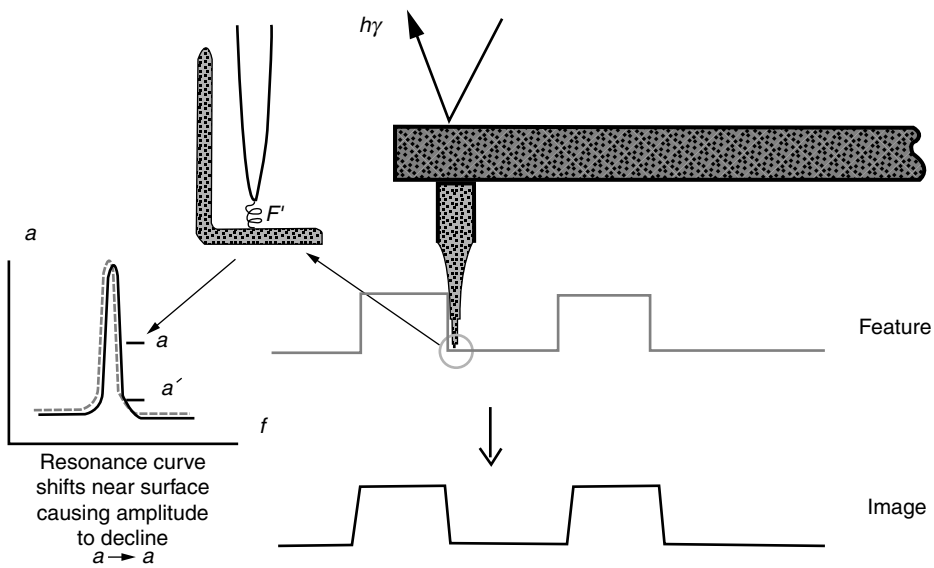


FIGURE 20.55 Principle of AFM scan. (From Rizvi, S., and A. Meyyappan, *Proc. SPIE*, 1999, 740–52.)

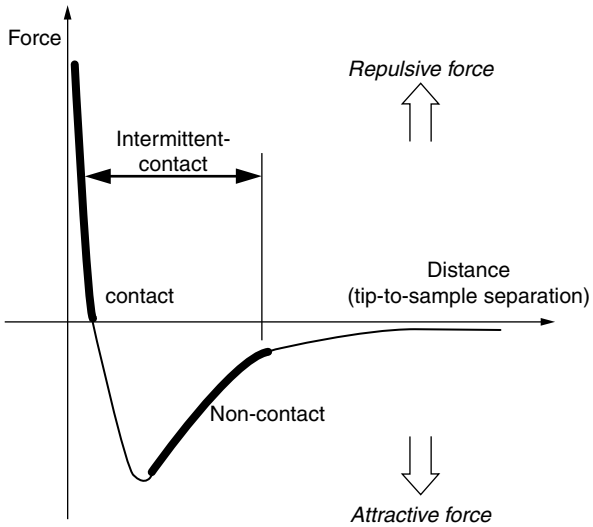


FIGURE 20.56 Inter-atomic force vs. distance. (From Muckenhirn, S., and A. Meyyappan, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 499–530. Boca Raton, FL, 2005.)

most common method to detect the position of the probe tip is by employing optical sensor. In one such scheme, a laser beam bounces of the back of the cantilever onto a position sensitive photo-detector [84,85]. During the scans the tip follows the contour of the surface which is monitored by a laser/detector system or any such precision motion detector system Scanning modes are referred as contact mode, alternative or intermittent contact mode, and non-contact mode.

In contact mode the probe and the sample surface are in soft contact. A typical sequence can be described as approaching the surface, contacting the surface, scanning the surface, retracting away from the surface. The force of interaction is in the range of 10^{-8}–10^{-6} N [86].

In an alternative or intermittent contact mode, popularly known as tapping mode, the tip is intermittently in contact with the surface of the sample. A typical sequence can be described as approaching the surface, contacting the surface, scanning the surface with alternate contact/retract of small vertical amplitude and high frequency, retracting away from the surface. Interaction forces: 10^{-9}–10^{-7} N [87].

In non-contact mode the probe is not in touch with the sample but is maintained at a small distance away from the sample by maintaining the attractive force at constant level. A typical sequence can be described as approaching the surface, sensing the surface, scanning over the surface at a defined flying height with small vertical amplitude, retracting away from the surface. Here the interaction forces are in the order of 10^{-12}–10^{-10} N [88].

20.5.1.1.3.1 Shape of Tips

Another area of interest is the shape of the tip of the probe. Once again this would depend upon the task on hand. A conical shaped tip can be useful for measuring CD profile and whereas a flare shaped tip can reach into undercuts inside a trench and could also give the information on the roughness of a trench wall. Figure 20.57 [82] shows the picture of a flare tip.

20.5.1.2 IP Metrology

20.5.1.2.1 Introducing IP

An important aspect of photomask metrology is the metrology of IP (or feature) placement. A pattern on a photomask represents the layout of a chip that is to be imaged on wafer, but the final chip ends up with

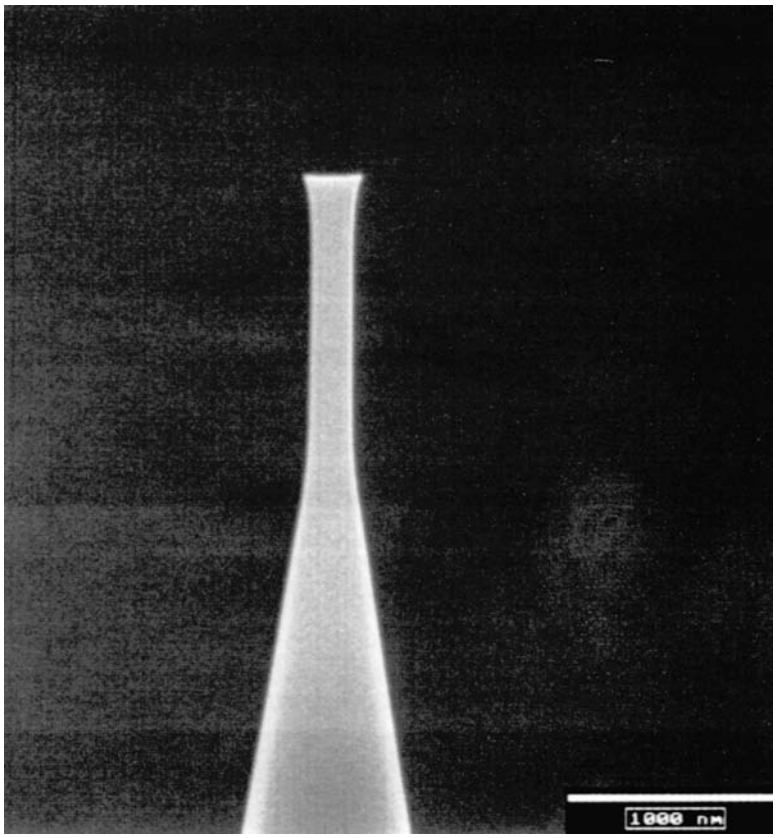


FIGURE 20.57 A flare-shaped tip. (From Rizvi, S., and A. Meyyappan, *Proc. SPIE*, 1999, 740–52.)

many layers and in most cases each layer is drawn on a separate plate. In order for the circuit to work, all layers must match. This would mean that all masks representing those layers, must also match. The masks representing these layers to form a chip are called a set of masks, or simply mask-set. Once a mask set is complete it must be checked to make sure all layers match. In early days of semiconductor industry, this type matching was done by a tool known as optical comparator, also known as overlay machines. When comparing two masks one could not tell which one of the two were closer to the actual design lay-out. In most cases a standard was chosen arbitrarily and the rest of plates were then checked against this de-facto standard.

This approach worked for a while but with the advancement in the semiconductor the information on feature placement in a coordinate system defined by the design layout became necessary. This meant that some type of coordinate measurement machine be developed where positions of features could be directly measured and compared with design values.

The first such machine capable of measuring the coordinate of a given feature was known as linear dimensional analyzer (LDA) introduced in the mid 1970s by Boller & Chivens, a company in Pasadena, CA. The system employed finely graduated glass rulers placed along the x and y axes of the measuring instrument. The system repeatability was about a micron. There were only a few such machine built and the program was discontinued because of their cost and that the specifications of the machine were below customer expectations. Later on, in the 1980s, Nikon Inc. introduced its Nikon XY-2I system with improved measurement capabilities and specifications. This machine lasted several generations and was

widely used in the industry. In the 1990s its *x–y* coordinate measuring system MS-2000 which then evolved into LMS IPRO series.

20.5.1.2.2 Description of the Machine

Currently Leica Inc. has dominated the industry with LMS-IPRO systems. The latest in the series LMS IPRO3 was unveiled in early 2005. For *X*/*Y* stage it employs laser interferometer system with 0.3 nm resolution that gives it long-term 3-Sigma repeatability of 2.5 nm over the full coverage of 230 mm mask [89].

20.5.1.2.3 Calibration and Principle of X and Y Coordinate System

Metrology of coordinate measurement is conceptually more complex than for CD metrology. There is a fundamental difference between the two kinds of measurements. In the case of CD, the measurement is one-dimensional and restricts itself within a very small range defined by the width of the feature which may be in the order of nanometer, or micron at the most. Coordinate measurement, on the other hand, involves two dimensions, *x* and *y*, and are to be measured over very long distances that are in the range of 200–300 mm. Moreover, precision and accuracy required for coordinate measurement remains to be in the same scale as required in the case of CD measurement. This makes the task of coordinate measurements even more difficult. Other area of consideration is the possibility of distortion in the grid defined by the *x* and *y* scales of the system. Here distortion means any kind of departure from perfect orthogonal grid. This types of issues are not encountered in CD metrology.

20.5.1.2.3.1 Calibration and De-Facto Standards

Choice of standard for calibrating IP instrument is another important issue that needs to be addressed when working with IP metrology.

A two-dimensional measuring system needs to be qualified or calibrated against some artifact (or standard plate) but the artifact itself must be qualified by some other machine. To put it simply, we need an artifact to calibrate a machine, but we also need a machine to calibrate the artifact. This difficulty does no arise in CD measurement which requires a measurement in one single dimension for which measurement standards are easily available. Many such standards are available in nature. For example the wavelength of light of some specified frequency is often used as a standard of length.

One approach to circumvent this difficult was first conceived by Michael Raugh in his 1985 paper titled "Absolute 2D sub-micron metrology fore e-beam lithography: a theory of calibration with applications" [90]. Raugh laid a ground work for self calibration where the coordinate system of the grid on the plate and the coordinate system of the *XY* measurement instrument could be used to calibrate each other. The mathematical approach of this theory makes an extensive use of group theory and sets of measurement to do the calibration. Since then there have been a number of studies to explore and make this theory applicable to mask measuring systems [91–95].

20.5.1.2.3.2 Addressing the Coordinate Systems

Initially, when a mask is written, it is written by a mask writer which has a coordinate system of its own and can be called coordinate system of the mask writer. The task of the mask writer is to reproduce the layout on the mask exactly as prescribed by its design rule. Hence at this point we are dealing with two coordinate systems, one belongs to the design and the other coordinate belongs to the mask writer. The design coordinate system is theoretical in its nature and is regarded as the absolute coordinate system. The coordinate system of the writer, since it relates to an actual machine, can be made to be infinitesimally close to the theoretical coordinate system but not likely to be perfect. Any departure from the theoretical system is regarded as the signature of the writer. The coordinate measuring machine for the same reasons can have signature of its own.

Hence one is working with three coordinate systems, theoretical or design, the coordinate system of the mask writer which is reflected on the plates and the coordinate system of the measuring instruments. And understanding the correct relationship among them is extremely important.

20.5.2 Inspection and Repair

20.5.2.1 Inspection of Defects

During its fabrication, the mask goes through various processing steps and no matter how controlled a process is, variation in process do happen. For example, at some places there may be chrome etched out which shouldnot have been, or on the other hand at some place chrome did not get etched where it's supposed to have been. These chrome related defects are called hard defects because they are not easily removable by any simple cleanup process. Such defects, if they are small, can be repaired either by selective removal of chrome or by selective deposition of opaque material as needed. Soft defects on the other hand are mainly resist residuals, stains, or water marks and are removable by some sort of a cleanup processes.

Defects are most critical when they happen to be on the pattern side of mask, since this surface is maintained at a perfect focus during the printing of mask on wafer.

As shown in Figure 20.58 [96], these defects can be on the chrome edge (defect 1a), on the glass (defect 1b), it can be a small transmission defect in the quartz surface, which is caused by the different quartz thickness (defect 1c); defects can also occur within the bulk of the glass surface (defect 1d). Also shown in the same figure are defects on the back side of mask (defect 2) that being slightly out of focus are more relaxed in their specification and not likely to be printed. However, defects that not necessarily get printed on wafer may still cause some sort of distortion in the pattern due to reduced intensity of light going through that portion of the mask.

20.5.2.1.1 Types of Defects

Some of the common types of hard defects are shown in Figure 20.59 [96] as chrome extensions (1), missing chrome (2), corner defects (3 and 4), pin holes (5) that are regarded as missing features, and pin dots (6) that are regarded as added features. Transmission defects in opaque region (7), and appearance of semi-transparent films in the clear region (8) also fall under the category of hard defect. Moreover, defects such as scratches on mask surface or bubbles, or any such defect in the bulk of glass also can be seen as a hard defect.

There are also hard defects that relate to feature misplacements and missizing as shown in Figure 20.60 [96]. Feature misplacement can be caused by errors in the original data preparation of a mask, as well as by mask writer errors. Related to feature there can be a sizing error. A feature can be mis-sized in either direction x or y.

Another type of hard defects is global CD and/or quality change over the entire mask. Edge roughness is one example, as well as global CD uniformity changes due to heating or etching effects. As these changes usually occur gradually over the span of the mask area, state-of the art mask inspection hardly detects them.

20.5.2.1.1.1 PSM Related Defects

In phase shift masks there can be defects that can be phase specific defects. Figure 20.61 [96] shows some of the phase shift related defects in Half Tone PSMs. A mask might have additional phase material (defect

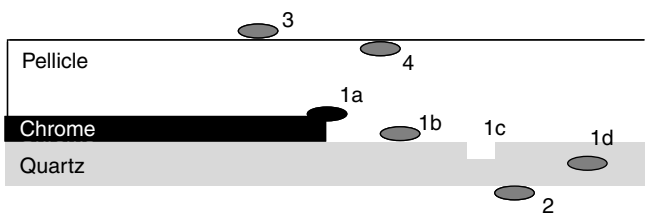


FIGURE 20.58 Defects of four different surfaces of mask. (From Rosenbusch, A., and S. Hemar, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 589–98. Boca Raton, FL, 2005.)

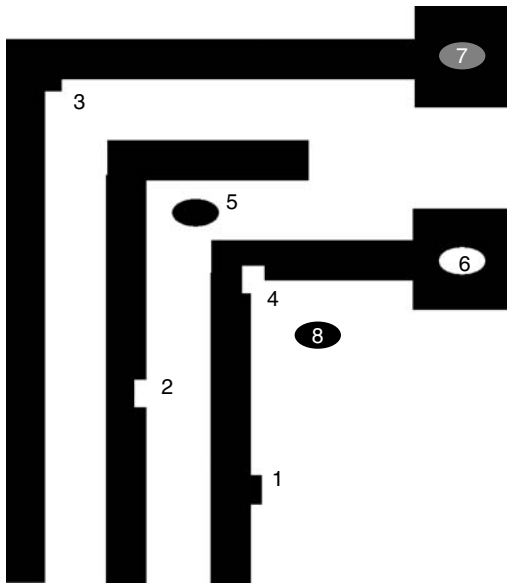


FIGURE 20.59 Hard defects on COG mask. (From Rosenbusch, A., and S. Hemar, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 589–98. Boca Raton, FL, 2005.)

1 and defect 2) or missing shifter material (defect 3). In HT.PSM absorber material is slightly transmitive and hence a shifter defect is also transmitive to an extent. This poses a challenge to mask inspection systems since the conventional inspection schemes are designed for binary masks that are based upon 100 percent transmission or zero transmission only.

Besides HTPSMs there are Alt.PSMs where the defect can occur during the manufacturing process involved in generating an Alt.PSM. In an Alt.PSM the shifter area is generated, for example, by an additional etch step. This step might produce defect as shown in Figure 20.62 [96]. The defect 1 is an

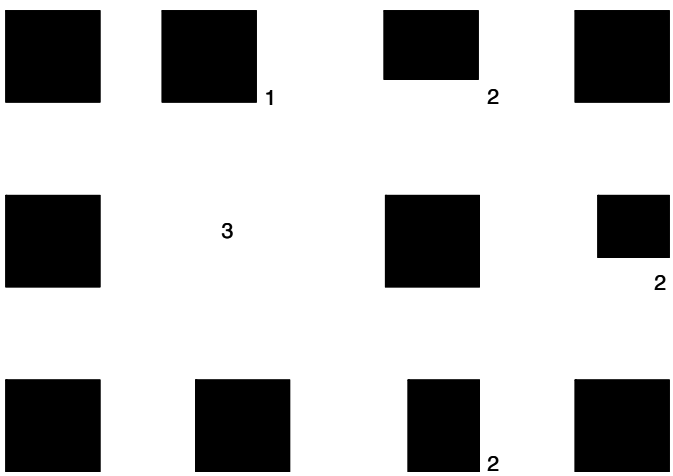


FIGURE 20.60 Example of feature misplacement. (From Rosenbusch, A., and S. Hemar, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 589–98. Boca Raton, FL, 2005.)

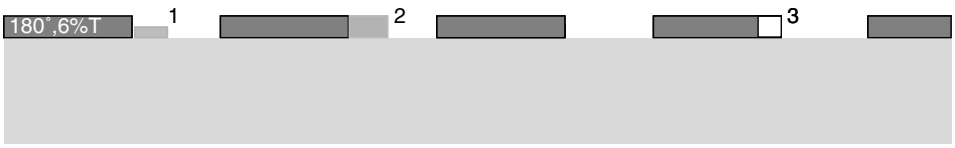


FIGURE 20.61 Phase related defects on halftone PSM (HTPSM). (From Rosenbusch, A., and S. Hemar, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 589–98. Boca Raton, FL, 2005.)

FIGURE 20.62 Phase related defects on Alt.PSM. (From Rosenbusch, A., and S. Hemar, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 589–98. Boca Raton, FL, 2005.)

additional partial shifter in the phase area. This defect might change the phase behavior of the mask. The defect 2 presents additional full shifter, hence the target CD (generated by the interference of non- and shifter area) might not be guaranteed by the mask. The defects 3 and 4 are more conventional. Additional absorber is located in the shifter or non-shifter area. The inspection support of these types of mask poses the biggest challenge for mask inspection. Phase errors might only be seen at the exposure tool wavelength.

20.5.2.1.1.2 Minimum Defects Requirement

At each generation of semiconductor lithography a minimum defect size, based on minimum gate width has been defined. These are listed in the SIA roadmap. Table 20.4 [96] shows the minimum defect requirement as defined in the latest international technology roadmap for semiconductor (ITRS), 2003. It is becoming customary to define minimum defect size in terms of not only its printability but also on its impact on CD variability for which aerial imaging (to be discussed later) has to be employed.

20.5.2.1.2 Basic Principle of Mask Inspection

There are two basic principles for inspecting masks

1. Die to database inspection
2. Die to die inspection

The first method compares the printed mask feature to their actual designs. A die-to-database inspection will detect all differences between the database and mask itself. If a difference violates the defect criteria given, it will then be classified as a defect.

TABLE 20.4 Minimum Defect Size Criteria as Defined by ITRS 2003

Year	2003	2004	2005	2006	2007	2008
Minimum defect size (nm)	80	72	64	56	52	45.6

Source: From Rosenbusch, A,. and S. Hemar, In *Handbook of Photomask Manufacturing Technology*, ed. Rizvi, S., Taylor & Francis Group, 589–98. Boca Raton, FL, 2005.

FIGURE 20.63 Die to database inspection. (From Rosenbusch, A., and S. Hemar, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 589–98. Boca Raton, FL, 2005.)

Figure 20.63 [96] shows a die-to-database comparison of a gate feature. On the left is the database image. The feature is well defined with straight edges and sharp corners. The image on the right side shows the image of the mask feature. The challenge of die-to-database inspection is to identify real mask defects distinguishing them from the systematic errors like edge roughness (etch step) or corner rounding (mask writer) introduced by the mask manufacturing itself.

If the mask has more than one die, a second method, known as die-to-die inspection method can be applied. This method assumes that all dies of a mask are similar. Die-to-die inspection between all dies will identify all defects. The drawback of this method is that it cannot identify a systematic error that can occur in all dies, such as additional feature or a little extension as shown in the bottom of Figure 20.64 [96].

20.5.2.2 Repair of Defects

Toward the completion of photomask there may still be some defects found during their final inspection. At this point it would be cost effective to repair them if they are small and repairable. Basically the defects can be opaque (extra chrome) or clear (missing chrome). In one case the extra material has to be removed while in the other case areas have to be patched up. In some cases such as in the case of HTPSM the task of repair is not quite that simple.

Currently there are four repair techniques with their own strengths and weaknesses in specific areas. These are

Laser repair
Focused ion beam (FIB) repair
AFM nano-machining repair
Electron beam mask repair

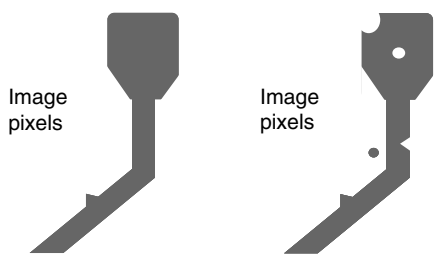


FIGURE 20.64 Die to die inspection. (From Rosenbusch, A., and S. Hemar, In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 589–98. Boca Raton, FL, 2005.)

20.5.2.2.1 Laser Repair

The use of laser beam to remove opaque defects has been applied since the early 1970s. The technique has proven to be highly reliable, easy to use, and has shown high throughput [97].

The lasers are pulsed with duration if the pulse can be in the order of nano, pico, or in femto-second scales. In case of nano and pico second lasers, the absorber heats up, melts, and evaporates by laser energy; this also can cause some splatter. There can also be some substrate damage creating phase related defects [98]. In the case of femtosecond laser the mechanism of absorber removal is somewhat different [99]. Here the laser pulse directly excites valence electrons to antibonding states causing the material to enter the vapor phase before the electrons transfer their energy to phonons (heat) and hence substrate damage is minimal. Repair quality of femtosecond laser tools is better than nano- and picosecond laser, but still limited by diffraction of laser light at the aperture used for defining the beam.

There are also laser tools for clear defects. Here, a gas is adsorbed locally on the surface. Laser energy heats up the substrate and gas, decomposing the gas and leaving a film behind the irradiated surface. Laser clear repair has not gained widespread market acceptance due to film qualities, size, and placement issues.

20.5.2.2.2 Focused Ion Beam Repair

In this case the repair machine utilizes focused ion beam (FIB) as probe that is also used for image formation.

The machine extracts gallium ions from a liquid metal ion source (LMIS) which is constantly replenished by a reservoir of liquid gallium. The gallium ions are accelerated down through the focusing optics of the column and strike the target where repairs are to be made.

The system is equipped with a scanning mechanism, with a set of electrostatic lenses and apertures where a beam can be used to form a spot and the spot can be moved to the position of interest (Figure 20.65) [100].

When gallium ion strikes the target it is implanted into the surface and the same time it also produces some localized sputtering in the neighborhood of its impact. The area of glass implanted with gallium ions becomes an absorber of photons and reduces the transmission [101]. This is called "gallium staining."

Clear repair is made by injecting specific gases into the area of interest where the ion beam is scanned. At this point gas molecules adsorbed on the surface can dissociate from the energy of the impinging ion. In some cases, materials will be deposited on the surface. Tungsten, carbon, platinum, and gold films (among others) have been deposited in such a fashion from precursor gases. Focused ion beam clear repair uses carbon as the deposition film of choice. Focused ion beam tools repair opaque defects by sputtering the film away, usually in the presence of a gas to enhance removal and selectivity [102]. MoSiON films for HTPSM [103] show high selectivity to the substrate with typical process gases, and very little substrate damage beneath the completed repairs [104]. Chrome, however, is very difficult to remove due to its crystalline form on the substrate and the fact that there are very few volatile chrome compounds. Physical sputtering is a major component of chrome removal. Sputtering rates increase at sharp topography edges. What this means is that during chrome removal, substrate damage is increased around the periphery of a defect and is known as "riverbed." Controlling quartz damage during FIB opaque repair is critical to success (Figure 20.66).

Carbon films have long been used to fix clear defects in FIB tools. The films display an excellent opacity and ruggedness in all cleaning processes, and FIB's are the preferred method for binary mask clear repair (Figure 20.67). These films do not have the same index of refraction as MoSiON films of EAPSM masks. Users can either match phase or transmission of EAPSM films, but not both at this time. Most users opt to match transmission. "Blob" defects can be repaired with FIB tools. Software overlays video images of defective and good mask areas to create a bitmap repair map. The operator then overlays the repair maps on the defect and starts the repair. The beam raster scans only over the repair area to fix these killer defects. This is known as pattern copy repair.

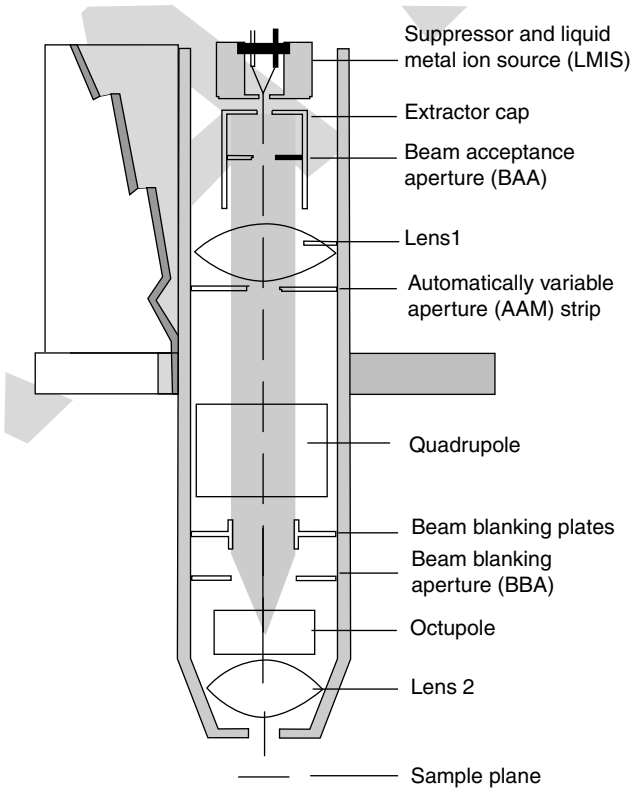


FIGURE 20.65 Schematic of focused ion beam column, courtesy FEI company. (From Lee, R., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 629–46. Boca Raton, FL, 2005.)

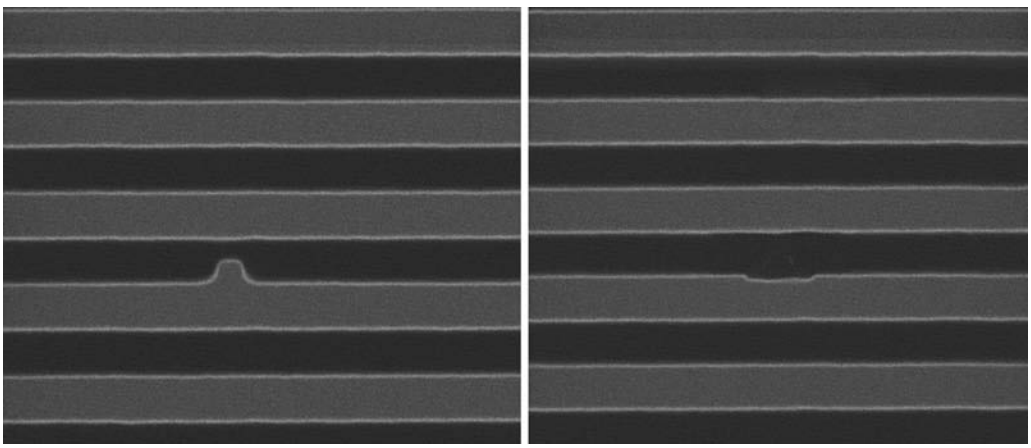


FIGURE 20.66 Chrome repair, before and after. Courtesy FEI company. (From Lee, R., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 629–46. Boca Raton, FL, 2005.)

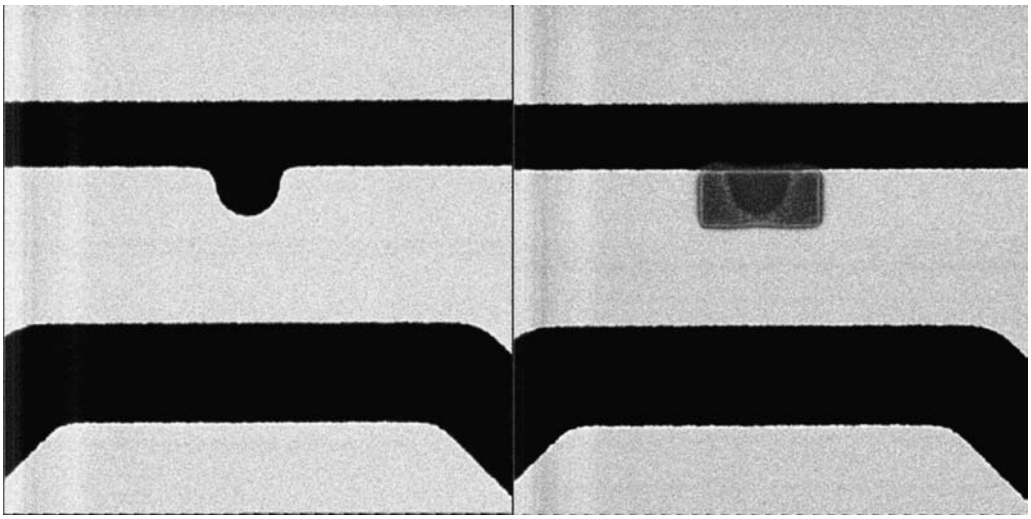


FIGURE 20.67 Clear repair, before and after. Courtesy FEI Company. (From Lee, R., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 629–46. Boca Raton, FL, 2005.)

20.5.2.2.3 AFM Nanomachining Mask Repair

In the past, AFMs have been used for the CD measurement or for reading the topography of some surface. Now new use of AFM has been envisioned that is called Nano-machining. In the case of nano-machining the tips not only makes actual contact with the surface but is used to abrade the defect through direct contact of the tip to the surface. [105,106] Special tip designs can machine down from the top surface, or cut in from the side on defects. The tool can image and remove opaque defects of chrome and MoSiON down very small sizes. Missing clear features have also been reconstructed in absorbers [107].

It has also been used to repair quartz defects and bumps on Alt PSM masks. (Figure 20.68) [99] Cryogenic carbon dioxide blown over the work area removes generated debris to keep the mask clean.

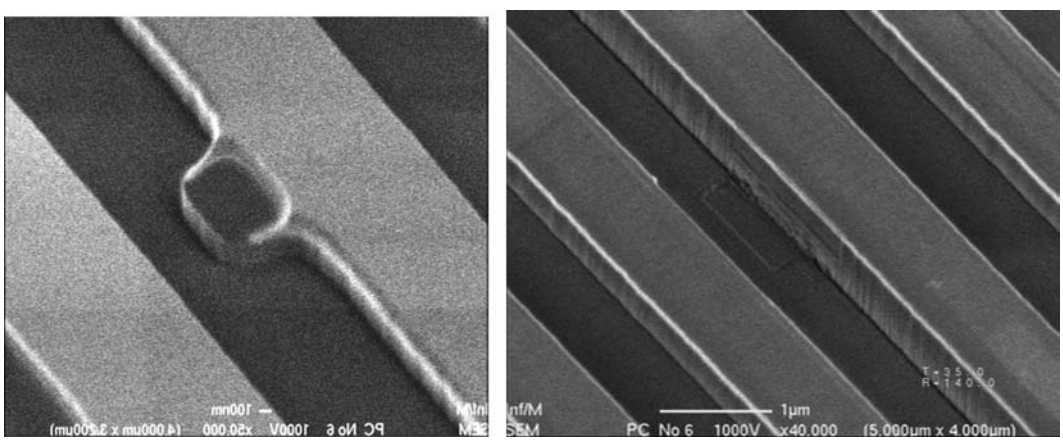


FIGURE 20.68 Quartz defect repair, before and after courtesy of RAVELLCFEI company. (From Lee, R., *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 629–46. Boca Raton, FL, 2005.)

20.5.2.2.3.1 Electron Beam Mask Repair

There have been some work reported in this area.

The standard electron beam for mask repair applications has a Gaussian profile. Electrons striking the surface either implant inside the substrate, drain off to ground through conductive paths, or escape as BSEs (essentially rebounding off atomic nuclei and leaving the substrate surface). Energy from momentum transfer of the impinging electrons also creates SEs that escape from the substrate. Both secondary and BSEs can be collected together or separately and used for image formation though they have different contrast mechanisms.

Many companies are actively pursuing e-beam chemistry research for a variety of applications, including mask repair [108]. Gases injected into the vacuum chamber and adsorbed on a substrate surface can be dissociated by the impinging electron beam, though the energy transfer by individual electrons at the gas has a much lower overall average value. Since there is no physical work such as sputtering (as in a FIB repair tool) at the substrate surface, the surface reaction becomes almost purely chemical from of interaction of the dissociated gas components and the target materials.

Some companies offer research tools for e-beam mask repair investigation [109]. Researchers performed seminar works in e-beam etch for tantalum nitride (TaN) and molybdenum Silicide (MoSiON) absorbers, and metal deposition for clear repair applications. Deposits have proved resistant to chemical attack from cleaning cycles, but must still be qualified for physical ruggedness from mechanical cleaning processes and examined for any issues with reflectivity. Chrome etch remains elusive for e-beam tools, however, as etch rates revealed to date are too slow for production mask repair. Companies promise commercial repair tools for the future and research proceeds swifty.

20.5.3 Phase, Transmission, and Image Evaluation

Phase shift mask is a type of mask that controls the phase of transmitted light. It is important that this shift in the phase be correct as prescribed by the designer of the PSM. Any deviation from the prescribed value not only degrades the resolution but also has an adverse effect on other aspects of imaging that are not easily recognizable such as focus offset, variation in pattern sizes, and others. For precise control on phase shift also requires means of measuring it. Since, a change in phase is also accompanied by a change transmission it is preferable that a machine designed for measuring phase shift should be capable of measuring transmission also. This is especially important for HTPSM where the shifter film also cause a change in transmission.

Capability of measuring phase and transmission adds a new dimension to the evaluation of mask that until a while back was limited to CD measurement only. The measured values of phase, transmission, or CD give information by which the image quality can be predicted but not observed until the image is printed on wafer or imaging parameters are run through some sort of modeling program. These can be time consuming and costly. However, in current year optical tools have been developed that can read a feature on a mask and provide a simulated image as it would be given by mask exposing system. This tool is known as aerial image measurement system tool, where AIMS is an acronym for and will be described in the later half of this section.

20.5.3.1 Phase and Transmission Measurement

There are two ways to measure the phase shift and transmittance through a PSM. One way is to directly measure the phase and transmittance using the same wavelength that would be used on the wafer exposure system. The other way is to measure the phase and transmittance using a different wavelength and do the calculation for the desired wavelength. The later approach, however, can give misleading results, especially in case of HT.PSM where shifting layer is non-uniform, composed of multilayer, or has some other uncertainities.

Transmittance addressed here refers to the transmittance of the film that excludes the contribution of form glass. Hence the transmittance here should be regarded as transmittance relative to glass. By the same token the transmittance of glass should be considered in reference to air.

20.5.3.1.1 Measurement Techniques

Direct measurement of phase shift and transmittance of PSM can now be made by MPM series of machines developed by Lasertec Corporation (Japan). [110] The MPM100 [111,112] uses the wavelength of 436 and 365 nm, whereas, their newest system MPM248 [113]used the wavelength of 248 and 193 nm. These machines are successfully being used in the industry across the globe. Lasertec Corporation has released MPM157 also that will be suitable for 157 nm lithography.

A schematic of MPM is shown in Figure 20.69. Light originating form its Hg–Xe source (for MP100 and MPM248) or Deuterium source (in case of 193 nm) transmits the mask pattern images on to an objective lens. The mask patterns are enlarged by the objective lens then laterally shifted by the Mach–Zehnder type image shearing interferometer which is placed behind the objective lens. After which the images are projected and overlapped at a pinhole and a camera. In the case of phase-shift measurement the images from the phase-shift pattern and from the non-shift pattern are made to overlap at the image plane which is same as the detector surface. These information among many others are employed for computing the phase shifts for the mask.

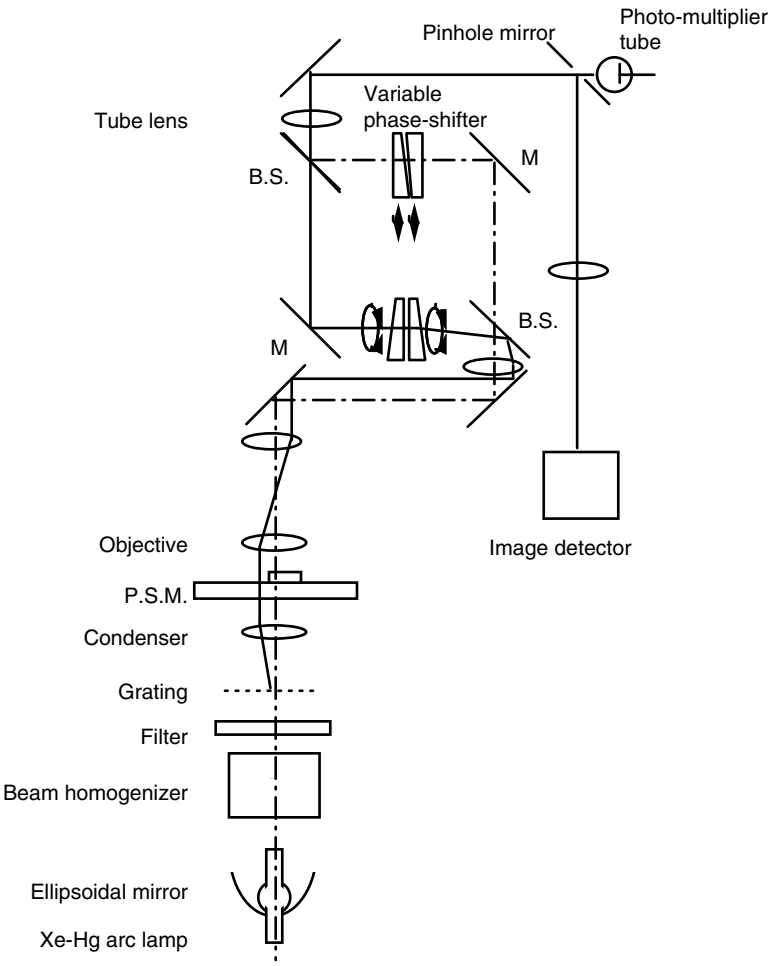


FIGURE 20.69 Optics for phase measurement. (From Kusunose, H., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 577–87. Boca Raton, FL, 2005.)

Transmittance by referencing quartz is defined as the ratio of the transmitted light intensities of two lights, one transmitted through shifter area and the other transmitted through non-shifter area.

There are a lot of unique technologies employed in the system, such as combination of low temporal coherent light illumination with interferometer, and periodically spatial coherent illumination. They are effective to improve lateral resolution and accuracy of the measurements. MPM series have been released with all wavelengths, which are commonly used in photolithography, and play a very important role to meet the requirement of PSM applications.

20.5.3.2 Image Evaluation

The photomask image to be printed on the wafer is known as the aerial image. Traditionally, image evaluations have been carried out by actually printing mask patterns on wafers under various conditions and then examining the images to select the set of conditions that can give the best image on wafer. Today's masks with OPC features and phase shift characteristics have become very complex and moreover today's exposure tools are also becoming more sophisticated especially in terms of the parameters that need to be tuned to print features with desired specification and quality. It thus has become apparent that the traditional way of image evaluation would not be cost effective and would lack precision needed today. For such reasons there has been an imminent need to develop a system that can do image evaluation in less time and be independent of operator's judgement.

Such a system is now available from Carl Zeiss and is known as AIMS™ [114]. This system optically emulates the image on a computer screen under various parameter settings [115–119]. Besides, looking at the image of features the system can also be used for evaluating the images of defects that a mask may carry.

20.5.3.2.1 System Description

Figure 20.70 shows the evolution of the Carl Zeiss AIMS™ tools for higher automation and shorter wavelengths. The two basic elements of AIMS™ tool are an illumination unit and an imaging unit. These

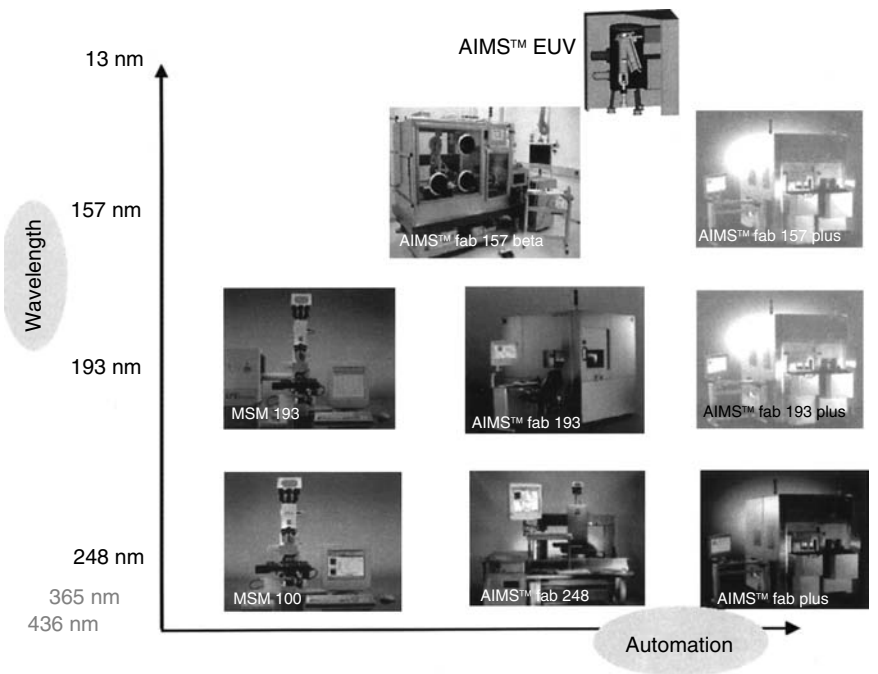


FIGURE 20.70 Evolution of Carl Zeiss tools. (From Zibold, A., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 607–28. Boca Raton, FL, 2005.)

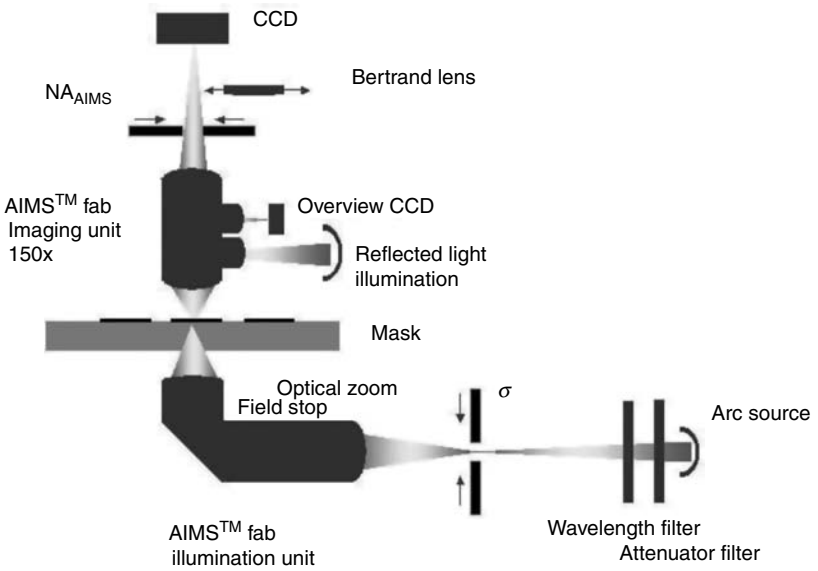


FIGURE 20.71 Optics of Carl Zeiss AIMS™ fabs platform. (From Zibold, A., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 607–28. Boca Raton, FL, 2005.)

are to replicate the conditions of an exposure tool. The illuminating system can be adjusted to give the required wavelength and the partial coherence (σ) whereas by arranging a set of pinholes in the right order can create an exposure tool equivalent of numerical aperture (NA). Apart from the wavelength settings the ranges for σ and NA are 0.25–1, and from 0.3 to 0.9, respectively, [120]. Figure 20.71 shows the schematic of the Carl Zeiss AIMS™ fab system for 248 and 365 nm exposure tool emulation. The optical set-up is a highly sophisticated derivative of the Carl Zeiss microlithography simulation microscope (MSM) based optics. Unlike the exposure systems which use reduction optics this system magnifies the mask image on the charge-coupled device screen. Typically a capture of the object of interest is made in a field of view of less than 60 μm×60 μm (standard 20 μm×20 μm).

20.5.3.2.2 Aerial Image Analysis

Aerial images acquired with the AIMS™ tools are recorded either as single images or as a through focus series (TFS) providing a stack of images. In Figure 20.72 an aerial image of a 5 μm wide isolated line on mask is shown. At best focus a strong effect of the edges can be seen due to interference and intensity values greater than unity. In the middle of the bright line the intensity is unity. Figure 20.73 shows a field of view from dense lines and spaces of pitch 1 μm 1:1 and a variety of standard plots which provide valuable information about the printability of mask defects. A quick prediction of the resist behavior can be obtained by the contour plot (lower right window). In a simple threshold model it is assumed that the expected resist linewidth depending on resist properties and exposure dose is given by selection of a threshold value, in this case the selected threshold value is 0.41. The intensity profile plot (upper middle window) allows the evaluation of peak intensities of bright features and comparison to one another. Further analysis with the Bossung plot can be displayed to extract information about the exposure latitude (lower left window) [121].

This machine can be very helpful in seeing how the different kinds of defects under given set of conditions can affect the features and how by repairing defect, the quality of the feature can be improved. Figure 20.74 gives an example of partially repaired phase defect and its quick visual evaluation of the printability by using a contour map.

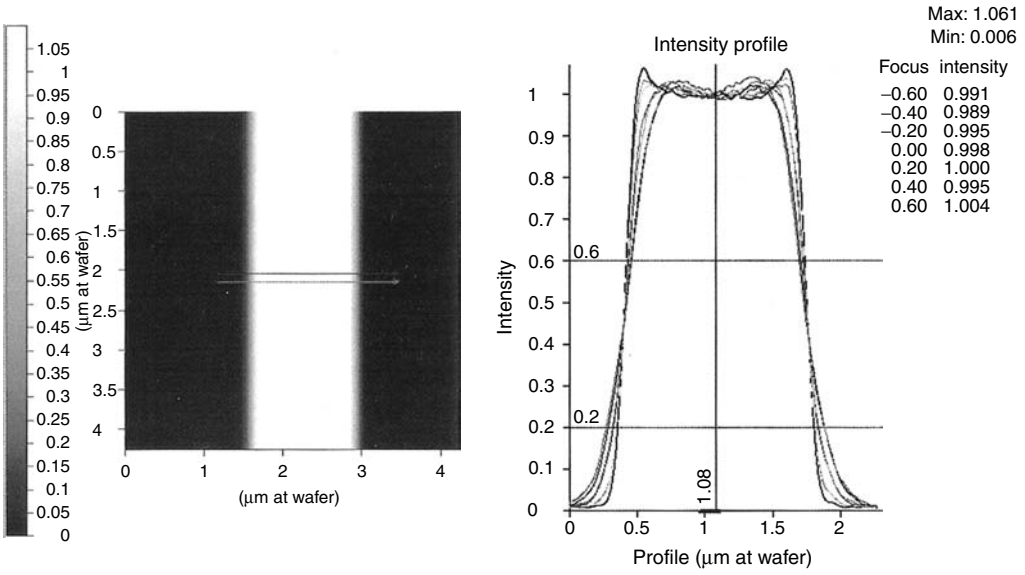


FIGURE 20.72 Normalized image of 5 μm line.

The latest system from Carl Zeiss is AIMS™ fab 193i meets the stringent requirement for advanced photomask evaluation for the 65 nm node.

20.6 Manufacturability and COO

Photomask technology precedes photomask manufacturability and can do so by significant time intervals. The technical capability is dependent to a large extent on the availability of equipment and

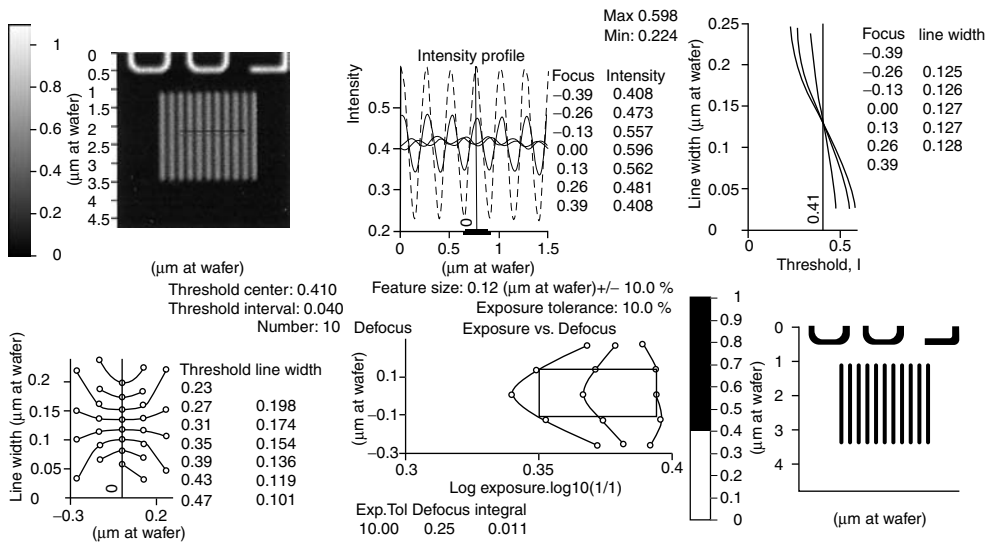


FIGURE 20.73 Various diagnostic types of information form AIMS™. (From Zibold, A., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 607–28. Boca Raton, FL, 2005.)

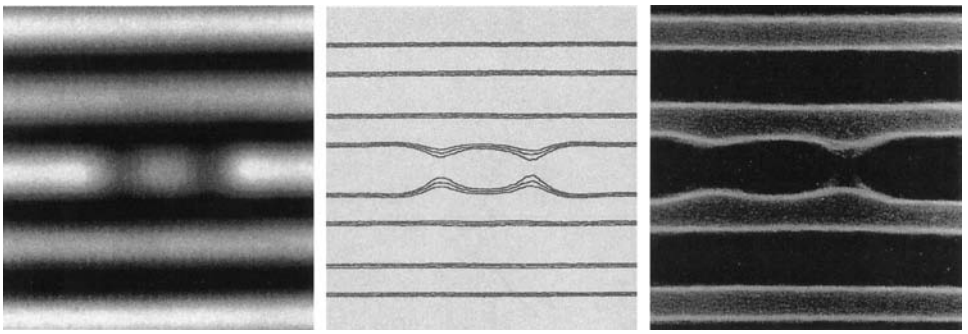


FIGURE 20.74 Aerial image (left) and printed feature of partially repaired defect. (From Zibold, A., In *Handbook of Photomask Manufacturing Technology*, Taylor & Francis Group, 607–28. Boca Raton, FL, 2005.)

materials (writer, inspection, resist, etc.). These tools and materials are typically available to the industry and the only limitation is capital cost and budgeting.

The load of photomask manufacturing is primarily carried out by the mask supplier. However, the initial development of any newly introduced, as always, remains partnership between the supplier and user.

20.6.1 Photomask Yield

High volume production of photomask manufacturing requires that all engineering issues be resolved earlier in the game [122–124]. This is a stage where all critical metrics be driven by a statistically maintained process and all manufacturing processes be steady except for minor tweaks. These manufacturing processes must also be under formal engineering control signoff procedures.

As Kurt Kimmel of IBM (formerly an assignee at International SEMATECH) says "Mask yield is the overwhelming driver for mask costs and delivery times"

Mask yield at mask supplier translates to the following differentiators for a mask customer:

Mask cycle time to mask user
On time delivery performance
Cost

20.6.2 Cycle Time to Mask User

Mask cycle time at the mask house translates directly into fab's ability to meet commitment to its customer. Even when a mask house at some point may run into an yield problem it can still meet fab's requirement by simply starting multiple number of plates to result into the required number of "good" plates.

Increased mask cycle times provide a fab with: (1) faster silicon learning curve if necessary, (2) reduced tool standby time waiting for mask, (3) increased fab tool utilization, and (4) quicker correction to, all of which result in meeting fab's commitment.

Design in/design wins are critical for semiconductor manufacturing. As chip customers finalize product design, chip manufacturers who deliver the first working proto-type get in the door first and can win all sales for that unit over a competitor who has not yet delivered a working chip.

20.6.3 On-Time Delivery Performance

Volume silicon manufacturing relies on consistent and planned wafer moves through a semiconductor fab. Any bottle-neck in the mask procurement process translates to a bottle-neck in silicon manufacturing.

Process tools sit idle while waiting for masks. Utilization drops. Bubbles build while awaiting process. Cost per wafer increases.

Consistency and reliability in on-time delivery ensures a smooth silicon manufacturing flow. While fast cycle time is important, once a technology node enters volume production, on time delivery becomes much more important than fast cycle time. Short cycle time with low on time delivery causes missed customer commitments, silicon inventory, and bottle-necks.

20.6.4 Cost

Mask cost is the final measure of mask manufacturing. If mask shop yield is high, mask shop manufacturing is monitored, reliable, and dependable. Higher yield must translate directly to lower cost and lower mask cost allows lower silicon cost. As mask costs are reduced, more designs can be released.

The number of wafers per design is a good indicator of mask usage. Franklin Kalk, CTO of Toppan Photomask, Inc. [125] states that number of designs has declined by approximately 13% per year since 1998, primarily due to a higher mask cost. The cost of building a plate since the 500 nm node to the current 65 nm node has also increased by 200×.

Often, cost has been a secondary factor for microprocessor and memory customers since it can be amortized over time, silicon, and customers. On the other hand, mask cost is a primary concern for logic customers. For newer technology nodes, mask costs are expected to be high enough that all types of mask users are concerned.

W. Trybula, Phil Seidel, and Ed Muzio of International SEMATECH have reviewed mask cost of ownership extensively. They have modeled mask cost looking at technology node, number of wafers per mask usage, mask yield, and plate type. IC Knowledge has completed analysis showing that mask cost consumes 56% of application-specific integrated circuit (ASIC), 10% of dynamic random access memory (DRAM), and 24% of logic devices at 130 nm node [126]. A mask-set cost is estimated to exceed $1 million dollars at this node.

Major contributors in rank order are believed to be data write time (optical or e-beam) and defect inspection by both TSMC and International SEMATECH [127].

20.6.5 Cost of Ownership

Mask cost of ownership is a difficult metric to calculate. Parameters vary from captive vs. merchant mask houses, type of product (ASIC, memory, logic), killer defects, wafer throughput and usage per mask set, mask inspectability, etc., [128–135]. Other factors which impact CoO are foundry, integrated device manufacturer (IDM), fab-less customer, use of zebra or shuttle masks, lithography strategy (such as 248 nm Alt.PSM vs. 193 nm binary, 4× vs. 5×), chip design (pattern density, die size, field size), and quantity and type of OPC. These factors add cost by increasing level of complexity to the mask manufacturing process and it is difficult to quantify them other than by relative comparison.

SEMATECH has modeled CoO over the years to a general level of usefulness. Actual figures on mask cost are companies' propriety and are not readily available to the outside world.

Because of the variation in CoO parameters across users and customers, mask cost per set is the true metric of concern. Within the mask customer, relative comparisons are made across technology nodes, type of design (logic, ASIC, memory, microprocessor), time, supplier, etc., but the only useful number to the mask customer is the actual mask cost per set per design type.

20.6.6 Device Specific Reticle Issues

The discussion about device specific mask concerns revolves around cost of ownership. As mentioned above, CoO is a difficult number to calculate since there are many variables which cannot be typically included in standard CoO modeling. If the CoO calculation is used at a high level, then it can be used as a relative measure. Only a small number of different design types are discussed below.

20.6.6.1 Memory Specific Issues

Memory (DRAM type) typically has a relatively small number of designs or devices, very high number of wafers, repetitive patterns, and static wafer processes per node. Patterns are multi-die and allow die-to-die inspection. Once the node is qualified, there are no major changes to the process or design until the next node. The most critical process steps for DRAM are typically during the front end when the gate and capacitor are built.

Because memory manufacturers run a small number of designs per fab, more process equipment such as scanners and steppers can be dedicated to specific process steps. Within process uniformity and repeatability can be maximized and on-mask requirements can be tighter because that mask will be used on a limited tool set.

The above situation translates to an environment where focus on mask quality is high but inspectability is also high because of repetitive patterns across the chip and multi-die designs. Lithography processes and therefore mask requirements are fixed over the period of time that DRAM node is manufactured.

The same mask set can be used over many wafers and over an extended period of time. This scenario allows the mask cost to be amortized over many customers and chips.

20.6.6.2 Microprocessor Specific Issues

Microprocessor devices behave similarly to DRAM type devices in terms of mask CoO. The number of designs per node is low, quality requirements are high, and lithography processes do not vary significantly during the life of the node. Again, the frot end is the most critical for microprocessors during which the gate is built.

One key difference from DRAM's is that microprocessors may be single chip patterns with larger and irregular die sizes, most often during the prototyping phase when inspection capability is lower. Die-to-database pattern inspection may have to be used, increasing inspection cycle time and therefore mask cycle time, and yield issues may be more difficult to resolve. Larger die sizes also mean that there are more likelihood for more killer defects within the same die.

Fab planning is simpler for DRAM's and microprocessors since there is little variation in timing or scheduling of pattern releases or number of wafers per customer. There are fewer opportunities for bubbles and bottle-necks in the manufacturing process. Mask cycle time and on-time delivery are less critical since the same mask set can be used for a long time period and there is a constant flow of material through the fab.

20.6.6.3 ASIC Design

ASIC chips typically have small number of wafers per design and there are many designs per node. There can be as few as 1 wafer per design or customer to many hundreds of wafers per design or customer. The requirement for a full mask set is independent of the number of wafers and the cost of each mask set cannot consistently be amortized over multiple customers or wafers. The life of a design is very short and the mask is unusable for any other design or customer.

Because the number of wafers processed may vary significantly across a design, fab planning is critical and mask on-time delivery with fast cycle time becomes more important. Fab cycle time and therefore mask cycle time directly impact delivery to customer and customer commitment.

Mask inspectability is more difficult than in the DRAM case because a device may include multiple modules with varying designs across the chip. The pattern is often multi-die which allows die-to-die inspection, reducing inspection cycle time and increasing inspectability in manufacturing status.

ASIC chips are very cost sensitive. Mask cost becomes a major issue. Lower cost options such as shuttle masks [136] and zebra masks may be used. Shuttle masks use multiple designs per reticle to maximize mask material and minimize manufacturing costs across multiple designs and customers. Zebra masks use multiple layers of the same device on a reticle to do the same thing.

20.6.6.4 Logic Design

Logic devices are similar to ASIC in terms of fab and mask planning. The constant release of new designs requires multiple priorities for both fab and mask supplier, high number of masks with short cycle times, and reliable on-time delivery performance. Time to customer and meeting customer commitment are critical. Number of wafers per logic design varies depending on the design itself.

Logic design is more cost sensitive than DRAM or microprocessor but not much as ASIC due to the higher number of wafers per mask.

20.6.6.5 Analog Design

Analog devices are very much different than all logic or memory or microprocessor designs. Often, leading analog wafer manufacturing processes are at least a node or two behind logic processes. This difference implies the use of older fabs, process equipment, and materials than for logic and other devices discussed above. Typically, issues such as DUV lithography or new process materials have matured and technical problems are few in the wafer manufacturing environment for the leading analog technology. Number of customers and wafers per design are high.

Analog designs have more levels per device which increases mask cost per set but the technical requirements (e.g., specifications) are looser for leading analog designs than for leading logic designs. This allows mask manufacturers to have achieved full manufacturing yield for the analog requirements and costs are at commodity levels.

Analog devices are extremely cost sensitive and focus is on mask set cost reduction from the wafer manufacturing perspective due to the commodity nature of this mask technology requirement.

References

1. Yoshioka, N. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 135–56. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 6.
2. Maurer, W., and F. Schellenberg. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 163–89. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 8.
3. Mack, C. "The Natural Resolution." *Microlith. World* 8, no. 1 (1998): 10–1.
4. Wong, A. K. K. *Resolution Enhancement Techniques in Optical Lithography*. Bellingham, WA: SPIE Press, 2001.
5. Maurer, W., K. Satoh, and D. Samuels Fischer. "Pattern Transfer at $k_1 = 0.5$: Get 0.25 μm Lithography Ready for Manufacturing." *Proc. SPIE* 2726 (1996): 113–24.
6. Pierrat, C., A. Wong, and S. Vaidya. "Phase-Shifting Mask Topography Effects on Lithographic Image Quality." In *International Electron Devices Meeting (IEDM) Technical Digest*, 53–6. San Francisco, CA, 1992.
7. Wojcik, G., J. Mould Jr., R. Ferguson, R. Martino, and K. K. Low. "Some Image Modeling Issues for I-Line, 5× Phase Shifting Masks." *Proc. SPIE* 2197 (1994): 455–65.
8. Liebmann, L., I. Graur, W. Liepold, J. Oberschmidt, D. O'Grady, and D. Rigaill. "Alternating Phase Shifted Mask for Logic Gate Levels, Design and Mask Manufacturing." *Proc. SPIE* 3679 (1999): 27–37.
9. Jinbo, H., and Y. Yamashita. "Improvement of Phase-Shifter Edge Line Mask Method." *Jpn. J. Appl. Phys.* 30 (1991): 2998–3003.
10. Toh, K., G. Dao, R. Singh, and H. Gaw. "Chromeless Phase Shifting Masks: A New Approach to Phase Shifting Masks." *Proc. SPIE* 1496 (1990): 27–53.
11. Watanabe, H., Y. Todokoro, Y. Hirai, and M. Inoue. "Transparent Phase Shifting Mask with Multistage Phase Shifter and Comb Shaped Shifter." *Proc. SPIE* 1463 (1991): 101–10.
12. van den Broeke, D., J. F. Chen, T. Laidig, S. Hsu, K. Wampler, R. Socha, and J. Petersen. "Complex 2D Pattern Lithography at Lambda/4 Resolution Using Chromeless Phase Lithography (CPL)." *Proc. SPIE* 4691 (2002): 196–214.

13. Torres, A., and W. Maurer. "Alternatives to Alternating Phase Shift Masks for 65 nm." *Proc. SPIE* 4889 (2002): 540–50.
14. Duff, J. "Photomask Fabrication and Technology." *Course-at 25th Technology Conference BACUS Symposium*, Oct. 3–7, 2005.
15. Kalk, F. D., R. H. French, H. U. Alpay, and G. Hughes. "Attenuated Phase Shifting Photomask Fabricated from Cr-Based Embedded Shifter Blanks." *Proc. SPIE Int. Soc. Opt. Eng.* 2254 (1994): 64–70.
16. Ushida, M., H. Kobayashi, and K. Ueno. "Photomask Blank Quality and Functionality Improvement Challenges for the 130 nm Node and Below." *Yield Manage. Solut.* 3 (2000): 47–50.
17. Prouty, M., and A. Neureuther. "Optical Imaging with Phase Shifting Masks." *Proc. SPIE* 470 (1984): 228–32.
18. Terasawa, T., N. Hasegawa, T. Kurosaki, and T. Tanaka. "0.3 μm Optical Lithography Using a Phase Shifting Mask." *Proc. SPIE* 1088 (1989): 25–33.
19. Dolainsky, C., and W. Maurer. "Application of a Simple Resist Model to Fast Optical Proximity Correction." *Proc. SPIE* 3501 (1997): 480–774.
20. Maurer, W., C. Dolainsky, T. Waas, and H. Hartmann. "Proximity Correction in Optical Lithography by OPTISSIMO. GMM Fachbericht 21." 161–7. Berlin, Offenbach: VDE Verlag, 1997.
21. Saitou, N. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 59–98. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 4.
22. Satoh, H., Y. Nakayama, N. Saitou, and T. Kagami. "Silicon Shaping Mask for Electron Beam Cell Projection Lithography." *Proc. SPIE Int. Soc. Opt. Eng.* 2254 (1994): 122–3.
23. Crewe, A. V. "Some Space Charge Effects in Electron Probe Devices." *Optik* 52 (1978): 337–46.
24. Saitou, N. "Monte Carlo Simulation for the Energy Dissipation Profiles of 5–20 keV Electrons in Layered Structures." *Jpn. J. Appl. Phys.* 12, no. 6 (1973): 941–2.
25. Warkentin, P. A., and J. A. Schoeffel. "Scanning Laser Technology Applied to High Speed Reticle Writing." *Proc. SPIE* 633 (1986): 286–91.
26. Bohan, M. J., H. C. Hamaker, and W. Montgomery. "Implementation and Characterization of a DUV Raster-Scanned Mask Pattern Generation System." *Proc. SPIE Int. Soc. Opt. Eng.* 4562 (2002): 16–37.
27. Liden, P., T. Vikholm, L. Kjellberg, M. Bjuggren, K. Edgren, J. Larson, S. Haddleton, and P. Askebje. "CD Performance of A New High-Resolution Laser Pattern Generator." *Proc. SPIE Int. Soc. Opt. Eng.* 3873 (1999): 28–35 (Part 1–2).
28. Rydberg, C. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 99–132. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 5.
29. Ljungblad, U., T. Sandstrom, H. Buhre, P. Duerr, and H. Lakner. "New Architecture for Laser Pattern Generators for 130 nm and Beyond." *Proc. SPIE Int. Soc. Opt. Eng.* 4186 (2001): 16–21.
30. van Adrichem, P. M. J., and C. K. Kalus. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 19–42. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 2.
31. Saint, C., and J. Saint. In *IC Mask Design Essential Layout Technique.* 237–43. New York: McGraw Hill, 2002.
32. "An OASIS by the sea" *Solid State Technology Staff Report at 22nd Photomask Technology Conference BACUS Symposium*, Oct. 1–4, 2002.
33. Gesley, M., and M. A. McCord. "100 kV GHOST Electron Beam Proximity Correction on Tungsten X-Ray Masks." *J. Vac. Sci. Technol.* B12, no. 6 (1994): 3478–82.
34. Amdahl, G. M. "Validity of Single-Processor Approach to Achieving Large-Scale Computing Capability." In *Proceedings of AFIPS Conference*, 483–85. Reston, VA, 1967.
35. Gustafson, G. L. "Reevaluating Amdahl's Law." *CACM* 31, no. 5 (1988): 532–3.
36. Skinner, J. G. "Photomask Fabrication for Today and Tomorrow. Short Course 122." In *SPIE Education Service Program*, 2001.
37. Asahi Glass Succeeds in Development and Mass-Production of Synthetic Quartz Photomask Substrate (QC-i) for ArF Liquid Immersion Lithography." http://www.agc.co.jp/english/news/2004/0623e.pdf (accessed on February 16, 2007).

38. Skinner, J. G., T. R. Grove, A. November, H. Pfeiffer, and R. Singh. In *Handbook of Microlithography, and Microfabrication*, edited by P. Rai-Choudhury, 377–474. Washington: SPIE Optical Engineering Press.
39. "Zerodur" http://www.us.schott.com/lithotec/english/prodcuts/Zerodur/Zerodur.html?PHPSESSID=cb74815cd7e44ed2e0deeade392fea1d (accessed on February 16, 2007).
40. Walton, R. "Photo Blanks for Advanced Lithography." *Solid State Technol.* 46, no. 10 (2003): 26–8.
41. Wang, B. B. "Residual Birefringence in Photomask Substrates." *J. Microlith. Microfab. Miscrosyst.* 1, no. 1 (2002): 43–8.
42. Rathsack, N., D. Medeiros, and C. G. Willson. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 325–39. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 15.
43. Abboud, F., K. Baik, V. Chakarian, D. M. Cole, R. L. Dean, M. A. Gesley, H. Gillman, et al. *Proc. SPIE Int. Soc. Opt. Eng.* 4754 (2002): 704–15.
44. Hattori, Y., M. Kiyoshi, A. Ken-ichi, Y. Takayuki, U. Satosh, M. Taiga, N. Eiji, et al. *Proc. SPIE Int. Soc. Opt. Eng.* 4754 (2002): 696–703.
45. Groves, T. R. *J. Vac. Sci. Technol. B* 14 (1996): 3839–44.
46. Babin, S. *J. Vac. Sci. Technol. B* 21, no. 1 (2003): 135–40.
47. Babin, S. *J. Vac. Sci. Technol. B* 15, no. 6 (1997): 2209–13.
48. Willson, C. G. In *Introduction to Microlithography*, In 2nd ed. Washington, DC: American Chemical Society, 1994, chap. 3.
49. Constantine, C., D. J. Johnson, R. J. Westerman, T. Coleman, T. Faure, and L. Dubuque. *Proc. SPIE Int. Soc. Opt. Eng.* 3236 (1997): 94–103.
50. Siew, Y. K., G. Sarkar, X. Hu, J. Hui, A. See, and C. T. Chua. *J. Electrochem. Soc.* 147 (2000): 335–9.
51. Namatsu, H., Y. Takahashi, K. Yamazaki, T. Yamaguchi, M. Nagase, and K. Kurihara. *J. Vac. Sci. Technol. B* 16, no. 1 (1998): 69–76.
52. Mancini, D. P., K. A. Gehoski, E. Ainley, K. J. Nordquist, D. J. Resnick, T. C. Bailey, S. V. Sreenivasan, J. G. Ekerdt, and C. G. Willson. *J. Vac. Sci. Technol. B* 20, no. 6 (2002): 2896–901.
53. Namatsu, H., Y. Watanabe, K. Yamazaki, T. Yamaguchi, M. Nagase, Y. Ono, A. Fujiwara, and S. Horiguchi. *J. Vac. Sci. Technol. B* 21, no. 1 (2003): 1–5.
54. Falco, C. M., and J. M. van Delft. *J. Vac. Sci. Technol. B* 20, no. 6 (2002): 2932–6.
55. Ito, H., C. G. Willson, and J. M. J. Frechet. In *Digest of Technical Papers of 1982 Symposium on VLSI Technology*, 86–7, 1982.
56. Ito, H., and C. G. Willson. In *Technical Papers of SPE Regional Technical Conference on Photopolymers*, 331–53, 1982.
57. Ito, H., and C. G. Willson, U.S. Patent 4,491,628, 1985.
58. Handa, H., S. Yamauchi, K. Hosono, and Y. Miyahara. "Dry Etching Technology of Cr Films to Produce Fine Pattern Reticles under 720 nm with ZEP-7000." *19th Annu. Symp. Photomask Technol.* 3873 (1999): 98–106.
59. Constantine, C., R. Westremann, and J. Plumhoff. "Plasma Etch of Binary Cr Mask." *19th Annu. Symp. Photomask Technol.* 3873 (1999): 93–7.
60. Buie, M. J., B. Stoehr, and Y. C. Huang. "Chrome Etch for <0.13 μm." *21st Annu. BACUS Symp. Photomask Technol.* 4562 (2001): 633–40.
61. Mueller, M., S. Komarov, and K. H. Baik. "Dry Etching of Chrome for Photomasks for 100 nm Technology Using CAR." *Photomask Jpn* (2002): 350–60.
62. "Photomask Etch for 65nm Node" http://eurosemi.eu.com/front-end/features-full.php?id=5674 (accessed on February 16, 2007).
63. Anderson, S. A., R. B. Anderson, M. J. Buie, M. Chnadrachood, J. S. Clevenger, Y. Lee, N. Sandlin, and J. Ding. "Optimization of a 65 nm Alternating Phase Shift Quartz Etch Process." *23rd Annu. BACUS Symp. Photomask Technol.* 5256 (2003): 66–75.
64. DeJule, R. "Trends in Wafer Cleaning" http://www.reed-electronics.com/semiconductor/article/CA163977 (accessed on February 16, 2007).
65. "The Radiance Process" http://www.radianceprocess.com/rad.html (accessed on February 16, 2007).

66. Novak, R., I. Kashkoush, and G. S. Chen. "Today's Binary and EAPSMs Need Advanced Mask Cleaning Methods." *Solid State Technol.* February (2004): 45–6.
67. Yen, Y.-T., G. B. Wang, and R. Heuser. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 395–410. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 19.
68. Yen, Y.-T. U.S. Patent 4,759,990, 1988.
69. Hoobler, R. J. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 433–55. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 21.
70. Postek, M. T. *SPIE Proc.* 480 (1984): 109–18.
71. Postek, M. T. In *Handbook of Charged Particle Optics*, edited by J. Orloft, 1064–75. New York: CRC Press, 1997.
72. Postek, M. T. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 457–97. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 22.
73. Postek, M. T., A. E. Vladár, and M. H. Bennett. *SPIE 22nd BACUS Symp. Photomask Technol.* 4489 (2002): 293–308.
74. Postek, M. T., and A. E. Vladár. In *Handbook of Silicon Semiconductor Metrology*, edited by A. Diebold, 295–333. New York: Marcel Dekker, 2000, chap. 14.
75. Postek, M. T., and R. D. Larrabee. In *Concise Encyclopedia of Semiconducting Materials and Related Technologies Advancements in Material Sciences and Engineering*, edited by S. Mahajan, and L. C. Kimerling, 176–84. UK: Pergamon, 1992.
76. Postek, M. T., and D. C. Joy. *NBS J. Res.* 92, no. 3 (1987): 205–28.
77. Ahmed, T., S-R. Chen, H. M. Naguib, T. A. Brunner, and S. M. Stuber. *Proc. SPIE* 775 (1987): 80–8.
78. Bennett, M. H. *SPIE Crit. Rev.* 52 (1993): 189–229.
79. Bennet, M. H., and G. E. Fuller. *Microbeam Anal.* (1986): 649–52.
80. Robb, R. F. *Proc. SPIE* 775 (1987): 89–97.
81. Postek, M. T. In *Handbook of Charged Particle Optics*, edited by J. Orloff, 363–99. New York: CRC Press, Inc., 1997.
82. Rizvi, S., and A. Meyyappan. "Atomic Force Microscopy: A Diagnostic Tool for Mask Making in the Coming Years." *Proc. SPIE* (1999): 740–52.
83. Muckenhirn, S., and A. Meyyappan. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 499–530. Boca Raton, Fl: Taylor & Francis Group, 2005, chap. 23.
84. Meyer, G., and N. M. Amer. "Novel Optical Approach to AFM." *Appl. Phys. Lett.* 53 (1988): 1045.
85. Alexander, S., L. Hellemans, O. Marti, J. Schneir, V. Eling, P. K. Hansma, M. Longuire, and J. Gurly. *J. Appl. Phys.* 65 (1989): 164.
86. Weisenhorn, A. L., P. K. Hansma, T. R. Albrecht, and C. F. Quate. "Forces in Atomic Force Microscopy in Air and Water." *Appl. Phys. Lett.* 54 (1989): 2651.
87. Vie, D., H. G. Hansma, C. B. Prater, J. Massie, L. Fukunaga, J. Gurley, and V. Elings. "Tappin Mode Atomic Force Microscope in Liquids." *Appl. Phys. Lett.* 64 (1994): 1738.
88. Martin, Y., C. C. Williams, and H. K. Wickramasinghe. "Atomic Force Microscope—Force Mapping and Profiling on a Sub 100 Å Scale." *J. Appl. Phys.* 61 (1987): 4723.
89. Schlueter, G., K.-D. Roeth, C. Blaesing-Bangert, M. Ferber. "Next Generation Mask Metrology Tool" *Proc. SPIE Int. Soc. Opt. Eng.* 4754 (2002): 758–68.
90. Raugh, M. R. "Absolute 2-D Sub-Micron Metrology for Electron-Beam Lithography: A Theory of Calibration with Applications." *Precis. Eng.* 7, no. 1 (1985): 3–13.
91. Raugh, M. R. *A Group-Theoretic Approach to Calibrating Very High Precision Measuring Instrument*, ICIAM Poster Session P-19. Washington, DC, 1991 (Raugh paper).
92. Raugh, M. R. "Two-Dimensional Self-Calibration: Role of Symmetry and Invariant Sets of Points." *J. Vac. Sci. Technol. B* 15, no. 6 (1997): 2139–45 (Rizvi Paper).
93. Takac, M., and H. Whitney. "Stage Cartesian Self-Calibration: A Second Method." *SPIE BACUS* (1998): 3546–9.
94. Raugh, M., and S. Rizvi. "Improving Overlay by Self-Calibration in Position Metrology." In *First International Workshop on Statistical Metrology, VLSI Symposium*, 1996.
95. Takac, M. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 457–798. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 22.

96. Rosenbusch, A., and S. Hemar. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 589–98. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 27.
97. Yoshino, Y., Y. Morishige, S. Watanabe, Y. Kyusho, A. Ueda, T. Haneda, and M. Oomiya. "High Accuracy Laser Mask Repair System LM700A." *Proc. SPIE* 4186 (2000): 663–9.
98. Yan, P., Q. Qian, J. McCall, J. Langston, Y. Ger, J. Cho, and B. Hainsey. "Effect of Laser Mask Repair Induced Residue and Quartz Damage in Sub-Half Micron DUV Wafer Process." *Proc. SPIE* 2621 (1995): 158.
99. Wagner, A., R. A. Haight, and P. Longo. "MARS2: An Advanced Femtosecond Laser Mask Repair Tool." *Proc. SPIE* 4889 (2002): 457–68.
100. Lee, R. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 629–46. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 30.
101. Morgan, J., and T. B. Morrison. "Advanced Reticle Repair." *Solid State Technol.* 43, no. 7 (2000): 195–201.
102. Hagiwara, R., A. Yasaka, K. Aita, O. Takaoka, Y. Koyama, T. Kozakai, T. Doi., et al. "Advanced FIB Mask Repair Technology for 100 nm/ArF Lithography." *Proc. SPIE* 4889 (2002): 1056–64.
103. Lin, B. J. "The Attenuated Phase Shift Mask." *Solid State Technol.* January (1992): 43–7.
104. Marotta, C., J. Lessing, J. Marshman, and M. Ramstein. "Repair and Imaging of 193 nm MoSiON Phase Shift Photomasks." *SPIE* 4562 (2002): 1161–71.
105. Laurence, M. R., and M. Nano. *Proc. SPIE* 4186 (2000): 670–3.
106. LoBianco, B., R. White, and T. Nawrocki. "Use of Nanomachining for 100 nm Mask Repair." *Proc. SPIE* 4889 (2002): 909–21.
107. Brinkley, D., R. Bozak, B. Chiu, C. Ly, V. Tolani, and R. White. "Investigation of Nanomachining as a Technique for Geometry Reconstruction." *Proc. SPIE* 4889 (2002): 232–40.
108. Boegli, V. A., H. W. P. Koops, M. Budach, K. Edinger, O. Hoinkis, B. Weyrauch, R. Becker, et al. "Electron Beam-Induced Processes and Their Applicability to Mask Repair." *Proc. SPIE* 4889 (2002): 283–92.
109. Koops, H. W., K. Edinger, V. Boegli, J. Bihr, and J. Greiser. "Electron Beam Mask Repair with Induced Reactions." In *19th European Mask Conference on Mask Technology for Integrated Circuits and Micro-Components*, 2003.
110. Kusunose, H. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 577–87. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 26.
111. Fujita, H., H. Sano, H. Kusunose, H. Takizawa, K. Miyazaki, N. Awamura, T. Ode, and D. Awamura. "Performance of i-/g-Line Phase-Shift Measurement System MPM100." *Proc. SPIE* 2793 (1996): 497–512.
112. Kusunose, H., A. Nakae, J. Miyazaki, N. Yoshioka, H. Morimoto, K. Murayama, and K. Tsukamoto. "Phase Measurement System with Transmitted UV Light for Phase-Shifting Mask Inspection." *Proc. SPIE* 2254 (1994): 294–301.
113. Kusunose, H., N. Awamura, H. Takizawa, K. Miyazaki, T. Ode, and D. Awamura. "Direct Phase-Shift Measurement with Transmitted Deep-UV Illumination." *Proc. SPIE* 2793 (1996): 251–60.
114. Zibold, A. In *Handbook of Photomask Manufacturing Technology*, edited by S. Rizvi, 607–28. Boca Raton, FL: Taylor & Francis Group, 2005, chap. 29.
115. Budd, R. A., D. B. Dove, J. L. Staples, R. M. Martino, R. A. Ferguson, and J. T. Weed. "Development and Application of a New Tool for Lithographic Mask Evaluation, the Stepper Equivalent Aerial Image Measurement System, AIMS." *IBM J. Res. Dev.* 41, no. 12 (1997): 119–28.
116. "Next Generation AIMS™ Mask Qualification System Launched" http://www.zeiss.com/C1256A770030BCE0/WebViewTopNewsAIIE/200630AF2964E7FFC12571F0001C87A0?OpenDocument (accessed on February 16, 2007).
117. Budd, R. A., J. Staples, and D. B. Dove. "A New Tool for Phase Shift Mask Evaluation, the Stepper Equivalent Aerial Image Measurement System AIMS™." *Proc. SPIE Int. Soc. Opt. Eng.* 2087 (1993): 162–71.
118. Budd, R. A., D. B. Dove, J. Staples, H. Nasse, and W. Ulrich. "A New Mask Evaluation Tool: The Microlithography Simulation Microscope Aerial Image Measuring System." *Proc. SPIE Int. Soc. Opt. Eng.* 2197 (1994): 530–40.

119. Martino, R., R. Ferguson, R. Budd, J. Staples, L. Liebermann, A. Molles, D. Dove, and J. Weed. "Application of the Aerial Image Measurement System AIMS™ to the Analysis of Binary Mask Imaging and Resolution Enhancement Techniques." *Proc. SPIE Int. Soc. Opt. Eng.* 2197 (1994): 573–84.
120. Carl Zeiss SMT: Next generation AIMS™ Mask Qualification System Launched http://www.smt.zeiss.com/de/press/releases.nsf/200630AF2964E7FFC12571F0001C87A0/$File/PL_076_06_AIMS45.rtf (accessed on February 16, 2007).
121. Wong, A. K.-K. *Resolution Enhancement Techniques in Optical Lithography*, Vol. TT47. Washington, DC: SPIE Press, 1999.
122. "SEMATECH Members Target Mask Costs" http://www.edn.com/article/CA305892/html?partner=enews (accessed on February 16, 2007).
123. Kimmel, K. P. "Photomask Supply Partnership to Optimize Cost and Value", http://www.future-fab.com/document.asp?d_id=1660 (accessed on February 16, 2007).
124. Segal, J., L. Milor, and Y. Peng. "Reduction Baseline Defect Density through Modeling Random Defect Limited Yield," http://www.micromagazine.com/grabber.php3?URL=http//www.micromagazine.com/archive/00/01/segal.html# (accessed on February 16, 2007).
125. "No Mask Maker's Holiday Seen in Next Generation Lithography" http://www.photomask.com/news/recent_news/pdf/050519_ConFab_Kalk.pdf (accessed on February 16, 2007).
126. "Performance Driven OPC for Mask Cost Reduction" http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/proceedings/dac/&toc=comp/proceedings/isqed/2005/2301/00/2301toc.xml&DOI=10.1109/ISQED.2005.93 (accessed on February 16, 2007).
127. John. and B. Lin. "Mask Cost and Cycle Time Reduction" http://www.sematech.org/meetings/archives/litho/mask/20011001/E_TSMC.PDF (accessed on February 16, 2007).
128. Gupta, P., A. Kahng, D. Sylvester, and J. Yang. "Performance Driven OPC for Mask Cost Reduction" http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/proceedings/dac/&toc=comp/proceedings/isqed/2005/2301/00/2301toc.xml&DOI=10.1109/ISQED.2005.93 (accessed on February 16, 2007).
129. Zhang, Y., R. Gray, S. Chou, B. Rockwell, G. Xiao, H. Kamberian, R. Cottle, A. Wolleben, and C. Progler. "Mask Cost Analysis via Write Time Estimation." *Proc. SPIE Int. Soc. Opt. Eng.* 5756 (2005): 313–19. http://www.photronics.com/templates/techpapers/spie05_abstract/spie05_abstract6.pdf (accessed on February 16, 2007).
130. Lewotsk, K. "Mask Projections," http://oemagazine.com/fromTheMagazine/feb04/specialfocus.html (accessed on February 16, 2007).
131. Micro Magazine, "Mask Projections". http://www.micromagazine.com/grabber.php3?URL=http://micromagazine.com/archive/03/10/mask.html# (accessed on February 16, 2007).
132. Kahng, A. B. "Bringing Down NRE" http://vlsicad.ucsd.edu/Publications/Columns/column7.pdf (accessed on February 16, 2007).
133. Lee, D. A., R. White, and B. J. Grenon. "Addressing Mask Costs," http://www.mentor.com/products/ic_nanometer_design/news/articles/mask_costs.cfm (accessed on February 16, 2007).
134. Cataldo, A. "High Mask Costs Seen Impeding Chip Prototypes," http://www.eetimes.com/conf/isscc/showArticle.jhtml?articleId=17408173&kc=3681 (accessed on February 16, 2007).
135. Pramanik, D., H. Kamberian, C. Progler, M. Sanie, D. Pinto, "Cost Effective Strategies for ASIC Mask," *Proc SPIE Int. Soc. Opt. Eng.*, 5043 (2003): 142–53.
136. Chen, S.-Y., and E. G. Lynn. "Effective Placement of Chips on a Shuttle Mask." *Proc. SPIE Int. Soc. Opt. Eng.* 5130 (2003): 681–88.

21

Plasma Etch

21.1 Introduction **21**-1

21.2 Technical Basics of Plasmas Relevant to Plasma Etching **21**-2
Introduction • Plasma Physics and Chemistry • Plasma Etching Tools • Plasma Etch Issues

21.3 65–90 nm CMOS Etch Process Modules **21**-18
Shallow Trench Isolation Etch • Gate Stack Etch • Substrate Contact Dielectric Etch • Damage Considerations • Back End of Line Etch Processes and Dual Inlaid Processing

21.4 The Next Generation—45–32 nm Technology Nodes **21**-46
Patterning Related Challenges • Device Roadmap Etch Challenges for 45–32 nm Technology Nodes • Front End of Line Etch Processes • Back End of Line Etch Processing and Ultra Low-κ Dielectrics

21.5 Nanotechnology—22 nm and Beyond **21**-57
Device Roadmap Challenges • Imaging Limitation and Etch Interactions • Extension of Existing (45–32 nm) Etch Technology Node • Process and Equipment Requirements • Challengers to Silicon beyond 22 nm

21.6 Modeling of Plasma Etching Processes **21**-63

References **21**-64

Peter L. G. Ventzek
Shahid Rauf
Terry Sparks
Freescale Semiconductor, Inc.

21.1 Introduction

Since its inception in 1992, the International Technology Roadmap for Semiconductors (ITRS) has provided insight into the challenges and directions for the industry.[1] Today's state-of-the-art complimentary metal-oxide semiconductor (CMOS) and memory technologies were developed to meet these challenges and, as we look forward, there is an even accelerating pace for the technology development to meet the requirements. Our discussion of semiconductor etch technology is therefore directly connected to this reference. Table 21.1 includes the product characteristics critical to patterning as defined by the roadmap. Plasma etching plays a central role to the enabling of patterning technology, so defined.

International Technology Roadmap for Semiconductors technology nodes are defined by the half pitch of the minimum critical dimension (CD) features and have a projected timeline. For the 2004–2007 timeframe, the hp90 and hp65 nm technology nodes are the current state-of-the-art for manufacturing and the technology development focus, respectively. To introduce the hp45–hp32 nm technology nodes in the 2007–2010 timeframe, new structures will be needed to meet the device performance requirements while pushing the limits of optical lithography. At the hp45–hp32 nodes, the back end of line (BEOL) electrical interconnect will become a dominating factor as the line resistance of the copper-barrier

TABLE 21.1 Key Lithography Requirements from the 2003 International Technology Roadmap for Semiconductors (ITRS)

Technology Node	130 nm (2001)	65 nm (2007)	22 nm (2016)
Dynamic random access memory (DRAM) contact in resist (nm)	165	80	30
DRAM contact after etch (nm)	150	70	21
DRAM overlay (nm, 3σ)	46	23	8.8
Microprocessor unit (MPU) half pitch (nm)	150	65	22
MPU gate in resist (nm)	90	35	13
MPU gate length after etch (nm)	65	25	9
MPU gate critical dimension (CD) control (nm, 3σ)	5.3	2.2	0.8
Line edge roughness (LER) (nm, 3σ)	7	2	0.7

Data for 130 nm is from the 2001 ITRS.
Source: From International Technology Roadmap for Semiconductors, Semiconductor Industry Association, 2003 Edition.

metallization begins to increase sharply while the devices needs lower R×C interconnect performance. Already, at 65 nm, the etch processes for interconnect impacts its physical and electrical properties. The coupling between process science and design will be even more closely coupled at 45 nm and below. At the hp22 nm node and beyond, the gate dimension of a planar device will be less than 10 nm and many define this as the threshold for the new technology, nanotechnology, where we will begin to fabricate structures at the atomic level. The introduction of nanotechnology will require features that will pass the foreseeable capabilities of optical lithography, devices operating within the quantum region and structures approaching the atomic level scaling. Here, new category of materials and process technologies may be needed to meet the device performance bringing with them a wide array of etch challenges.

This chapter takes a fundamental process science and modules based approach in describing the state-of-the-art in plasma etch technology. The chapter begins with an overview of the relevant plasma physics and chemistry. Then a snapshot of the role plasma etching plays in current CMOS technology is provided. For example, etch modules in the front end of line (FEOL) and BEOL are described, drawing on evolutions in the literature and technology to describe how the process deterministically impacts device fabrication. We then examine the role plasma etching is playing in conjunction with new patterning technologies such as immersion lithography to meet the challenges associated with scaling to 45 and 32 nm. Next, we look ahead to the fabrication of novel and nanoscale devices. The chapter closes with an overview of modeling of plasma etch processes.

21.2 Technical Basics of Plasmas Relevant to Plasma Etching

21.2.1 Introduction

A gaseous plasma is obtained by ionizing atoms or molecules in the gas, thereby creating a fluid containing ions, electrons, and neutral particles. Although some degree of ionization will occur in any gas under most circumstances, the term "plasma" technically refers to the state where charge density in the gas is large enough for (1) the gas to remain almost electrically neutral and (2) electric field generated by the ionized gas to shield out the influence of external electric fields. One can create plasmas of most gases or materials, and plasmas exist over a broad range of gas pressure and temperature. A certain class of plasmas, those that are partially ionized ($n_{\text{charged}} \ll n_{\text{neutral}}$), non-equilibrium ($T_{\text{electron}} \gg T_{\text{gas}}, T_{\text{ion}}$) and chemically reactive, have been found to be very useful for materials etching (i.e., selective removal of material from solid surfaces) in the semiconductor industry. These plasmas and their etching related applications are the main subject of this chapter.

In the simplest application of plasma based materials etching, plasma of appropriate gases is formed above the material undergoing etching and reactive radicals in the plasma react with the material surface. The gases are chosen such that either the reaction byproducts are volatile or they can be easily dislodged

through energetic ion bombardment. The resulting etch effluents are pumped away by the vacuum system. Steps involved in a typical etch process are illustrated in Figure 21.1. Assume that we want to etch certain regions of material A while retaining the integrity of underlying material B (Figure 21.1a). Using an appropriate lithography technique, a photoresist mask is first formed on the substrate surface with openings in areas where material A needs to be removed (Figure 21.1b). Plasma of the appropriate gases is then used to etch material A (Figure 21.1c). Gases in the plasma are chosen such that they rapidly etch material A and are relatively non-reactive with material B. The etch process therefore ends when all exposed material A is removed. Last step is the removal of the mask (Figure 21.1d), which often involves plasmas as well.

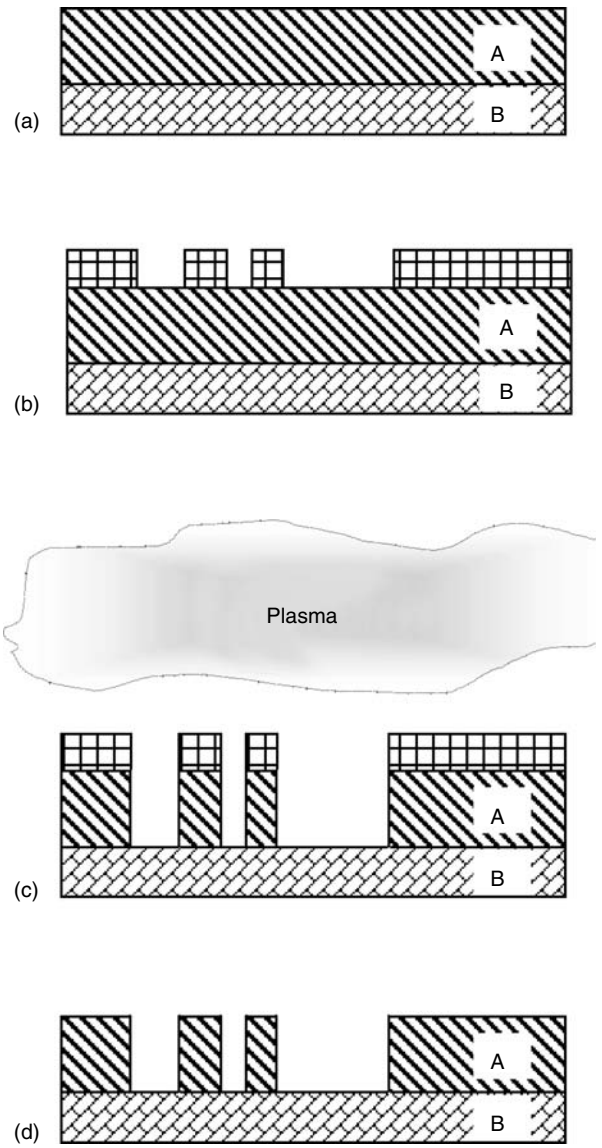


FIGURE 21.1 Steps involved in a typical patterning process: (a) film stack deposition, (b) photoresist deposition, imaging and development, (c) plasma etching, and (d) photoresist removal.

A wide variety of materials can be etched using plasmas and, over the last 35 years, many different gases and plasma reactor concepts have been utilized for materials etching. Later sections of this chapter describe etch processes of current and future interest for semiconductor manufacturing. The goal of this section is however to provide technical background of concepts and issues common to most etching applications. This includes basic properties of plasmas, chemistry of reactive plasmas, fundamental surface phenomena in plasma etching, common plasma etching tools, and engineering issues of relevance to plasma etch process design.

21.2.2 Plasma Physics and Chemistry

21.2.2.1 Plasma Generation and Basic Properties

A number of techniques can be used to ionize a gas to create a plasma. These techniques include radio frequency (RF) wave heating, laser irradiation, and gas heating. Most industrial plasma etching systems generate plasmas through electrical means, where the electrical frequency spans the range from DC to ultra high frequency (several GHz). A few stray electrons are almost always present in a gas, which have been generated by cosmic rays, thermal excitation or other means. When an electric field of sufficient strength is applied across the gas, these stray electrons accelerate, gain energy from the electric field, and bombard neutral atoms or molecules. If the energy of bombarding electrons is sufficiently large, they can ionize the impingent particles and generate additional electrons. This electron multiplication chain reaction can rapidly increase charged species density in the gas and create a plasma. Electron production within the gas is often complimented by secondary electron emission at material surfaces in commercial plasma etch systems. Secondary electron emission occurs under the influence of energetic ion bombardment on surfaces, and is the primary means of plasma production in DC plasma discharges. The primary electron loss mechanisms in plasma etchers include neutralization at the chamber walls and possibly electron attachment to molecules. Unless conditions are such that electron production rate is at least matched by the electron loss rate, a steady-state plasma cannot be sustained. This requirement defines the minimum RF electric field (i.e., RF power) and electron-neutral collision frequency range (i.e., gas pressure) where a stable plasma can be sustained in a given plasma system.

As was mentioned earlier, a charged gas mixture is only technically defined as plasma if the charged species density is large enough for the gas to exhibit certain collective electrical behavior. These include electro-neutrality in the bulk plasma and Debye shielding. Ions and electrons in the plasma generate an electric field. Once the charge density is large enough, this electric field does not allow positive and negative charges to separate and the gas remains electrically almost neutral. Furthermore, a plasma can shield out externally applied electric field (Debye shielding) and electric field within the bulk plasma is essentially due to charges within the plasma.

Most plasmas used for materials etching are not in thermal equilibrium. As a result, temperature of electrons ($T_{electron}$), which are the lightest component of the plasma, is substantially larger than neutral gas (T_{gas}) and ion temperature (T_{ion}). For typical commercial plasma etching systems, $T_{electron}$ is approximately in the range of 20,000–100,000 K, T_{ion} might be up to 2000 K, while neutral radicals and molecules are cooler than 1000 K. Electrons are therefore able to drive the chemistry within the gas at low gas temperatures. Electrons in the plasma are of course characterized by a broad energy distribution and the term "electron temperature" has only been used to characterize mean electron energy. Electron energy distribution is often approximated by the Maxwellian distribution ($\sim \exp[-\varepsilon/\kappa T_{electron}]$) in simplistic analyses, although the actual distribution is known to be quite non-Maxwellian. A detailed treatment of plasma properties can be found in the text by Lieberman and Lichtenberg.[2]

Two primary reasons that partially ionized non-equilibrium plasmas have found so many uses in industrial applications are (1) the presence of an electrical sheath at plasma-materials interface and (2) unique chemistry and reactive radical production at low gas temperature unachievable through other techniques. These two plasma characteristics are discussed in more detail in the next two sub-sections.

21.2.2.2 Sheath

In the preceding section, it was indicated that plasmas are electrically neutral in the central region of the chamber. However, as one approach the surfaces, electron, and ion densities start to diverge. This fundamentally happens because electrons are significantly lighter and therefore more mobile than positive ions. Without other restraints electrons will, therefore, tend to leave the plasma much faster than ions near the surface. The resulting deficiency of electrons near the surface generates an electric field whose direction is such as to retard electron departure.

The retarding electric field at the surfaces continues to grow in magnitude until positive and negative current through all the surfaces balance out. This region of intense electric field near plasma-material interface is known as sheath, and it plays a significant role in plasma etching processes.

Sheath potential and electrode currents in a parallel plate RF plasma are illustrated in Figure 21.2. As can be observed, the sheath potential is large and negative with respect to the surface for a large fraction of the RF cycle. During this time, current through the electrode is primarily due to positive ions or displacement current. For a brief period during each RF cycle, the magnitude of sheath potential decreases and electrons are able to flow out of the plasma discharge. Among other factors, sheath potential sensitively depends on how material surface is electrically connected. If the material is electrically floating on a dielectric, the initial electrons will charge material surface and the accumulating negative charge on the surface generates electric field that retards further electron migration to the surface. The resulting sheath potential is often referred to as the floating potential. When the material is electrically conducting and connected to other materials through the RF driving circuitry, balance of positive and negative current through all the electrode surfaces determines sheath potential at material interface with the plasma. If there are multiple electrodes in the RF circuit, the sheath potential at each individual surface will depend on the RF voltage applied to each electrode, the location of electrode with respect to plasma, area of the electrodes, and how electrodes are connected to the external RF driving circuit.

Anisotropic etching relies on bombardment of unidirectional energetic ions at the material surface. This is achieved in plasma etching reactors by accelerating positive ions in the sheath above the substrate surface. Since typical sheaths are thin ($\sim$10 μm–1 mm) and conformal with substrate surface, ion energy gain is primarily in the direction normal to the surface and the ion beam is essentially uni-directional there. For a given applied RF voltage, one would ideally like the largest possible fraction of the applied voltage to contribute to sheath potential at the substrate. Minimization of the sheath potential at other material surfaces also reduces ion bombardment induced damage to those components. Relative sheath potential drop at substrate surface can be maximized by careful reactor and RF circuit design. Some of the basic principles involved in circuit and reactor design are illustrated in Figure 21.3. Since positive and negative currents have to balance each other in the RF circuit, one technique that is often used to increase sheath voltage at powered electrode is to increase the relative surface area of grounded electrodes. When powered and grounded electrode areas are unequal, a circuit element, that helps further increase sheath voltage drop at the powered electrode, is a blocking capacitor (C_B) in series with the electrode. This blocking capacitor stores excess charge from the plasma, which increase sheath voltage at powered electrode.

Sheath potential and ion energy distribution function (IEDF) at substrate surface are also influenced by RF frequency and RF voltage waveform.[3] If RF bias frequency is low (typically less than 2 MHz), ion transit time through the sheath is less than the RF time period. Ion energy at the substrate is therefore determined by the instantaneous sheath electric field and, when averaged over time, the IEDF is broad. On the other hand, if the RF frequency is high (typically greater than 15 MHz), ions experience multiple cycles of the RF sheath potential and ion transit time through the sheath is large compared to the RF time. The resulting IEDF is therefore representative of the time averaged sheath potential and IEDF is narrow. The waveform of applied RF voltage also influences sheath voltage waveform, and this dependence is being used to help shape IEDF and therefore etching characteristics at the substrate.[4]

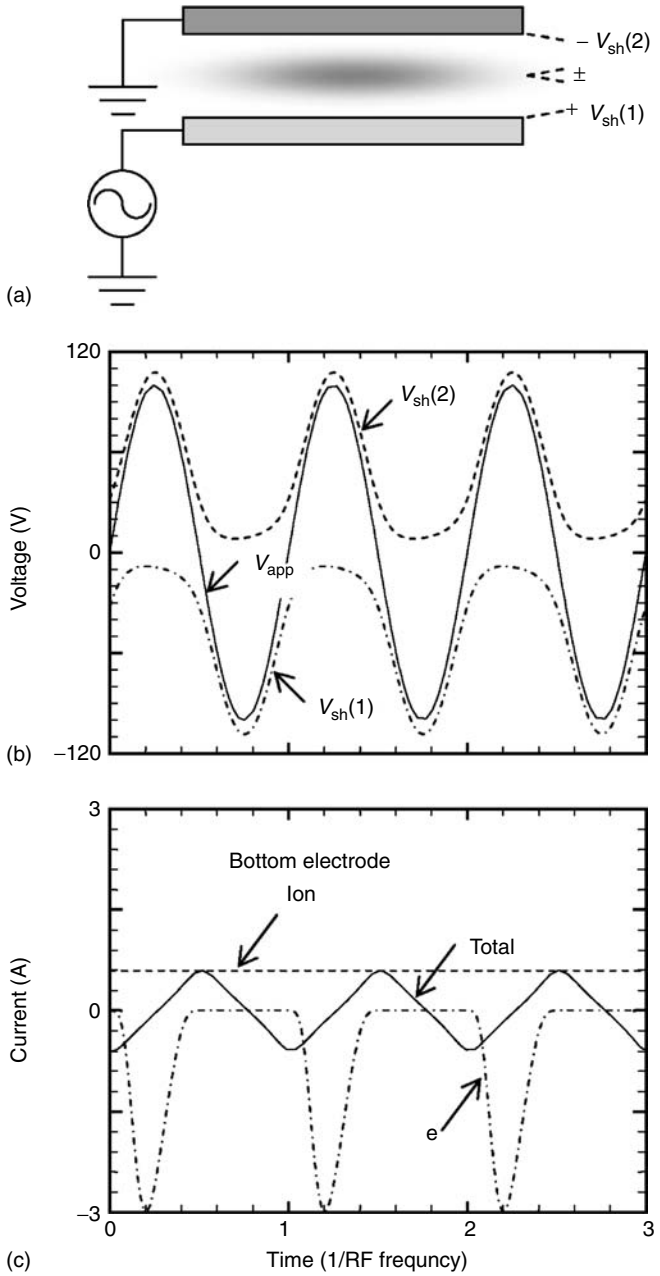


FIGURE 21.2 Sheath voltage and current in a parallel plate plasma reactor: (a) the parallel plate reactor, (b) applied and sheath voltages, and (c) ion, electron, and total current at bottom electrode.

21.2.2.3 Chemistry of Partially Ionized Plasmas

One unique feature of partially ionized plasmas is that electrons are extremely energetic even though the gas is at sufficiently low temperature (typically less than 1000 K). These energetic electrons can generate reactive radicals and ions, and drive chemical reactions that are unachievable through other means. An understanding of plasma chemistry and fundamental atomic and molecular processes in plasmas is

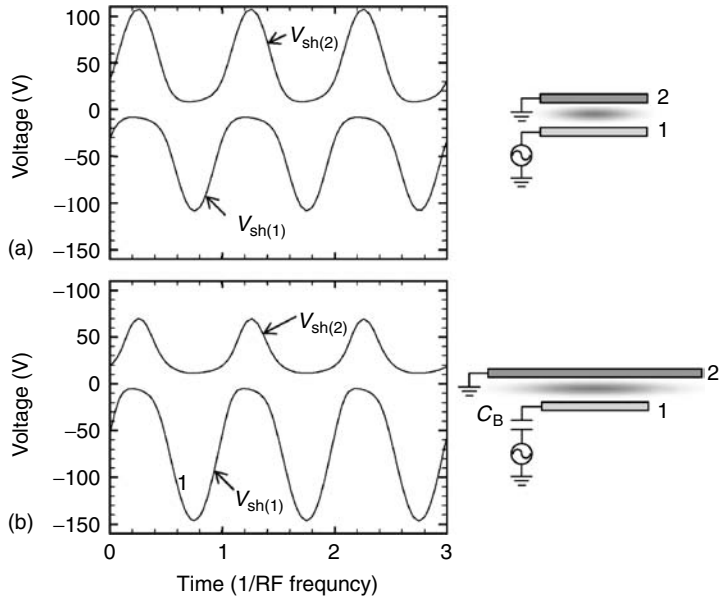


FIGURE 21.3 Sheath voltages in a parallel plate plasma reactor: (a) equal electrode areas and (b) unequal electrode areas with blocking capacitor.

invaluable for etch process engineering.[5] The fundamental chemical mechanisms in plasmas are reviewed in this sub-section.

Electrons drive many of the important gas phase processes in plasmas, especially in low gas pressure plasma etch systems in common use today. The two primary gas phase processes that generate plasmas are electron impact ionization and dissociative ionization. Some typical ionization reactions are shown below:

$$e + Ar \rightarrow Ar^+ + 2e \tag{21.1}$$

$$e + O_2 \rightarrow O + O^+ + 2e \tag{21.2}$$

$$e + CF_4 \rightarrow F + CF_3^+ + 2e \tag{21.3}$$

Ionization requires that outer shell electrons move far enough away from the parent atom or molecule that they are no longer subject to their attractive force. Ionization processes are therefore characterized by a threshold electron energy, below which ionization does not occur. While simple stripping of outer shell electrons is the only ionization phenomenon that is important for atoms in etching-relevant plasmas, more complex molecules and radicals can dissociate into two or more fragments during ionization.

In between the ground and ionized states, outer-shell electrons can be excited to a multitude of levels. Some of these levels are metastable while others are transitory. A substantial fraction of the input RF power is consumed by these excitation processes in etching-relevant plasmas. Metastable atomic states (e.g., Ar*) can build up to sufficiently large densities in plasma and, by virtue of their energy, they play an active role in driving plasma chemistry. In addition to electronic excitations (i.e., excitation of electrons to higher energy states), energetic electrons can also lead to vibrational and rotational excitations in molecules and radicals. A large fraction of the energy consumed in electron excitations leaves the plasma system as radiation. Depending on energy of transition states, this radiation can span the range from microwave to ultraviolet in etching relevant plasmas. Radiation from plasmas is finding many applications for plasma diagnostics and control, such as optical end point detection of etching processes.

Stable etchant molecules (e.g., CF_4, Cl_2) often do not react with semiconductor materials at room temperature. Electron impact dissociation is one fundamental molecular process that helps generate

chemically reactive neutral radicals in the plasma. Energetic electrons break up parent molecules into two or more neutral fragments during dissociation. For example,

$$e + O_2 \rightarrow O + O + e \tag{21.4}$$

$$e + CF_4 \rightarrow F + CF_3 + e \tag{21.5}$$

$$e + Cl_2 \rightarrow Cl + Cl + e \tag{21.6}$$

Radicals produced during dissociation tend to be more reactive than the parent gases, and these radicals help drive the surface processes and plasma chemistry. Similar to ionization, the dissociation processes have a threshold electron energy below which no dissociation occurs. The dissociation thresholds however tend to be smaller than the ionization thresholds, resulting in dissociation rates being much larger than ionization rates under many conditions. Many plasma processing gases therefore almost completely fragment in high density etch plasma systems, while the ionization fraction (n_{ion}/n_{gas}) rarely exceeds 10^{-3}.

If the gas pressure is sufficiently low (typically less than 15 mTorr), electron driven processes primarily determine what goes on in the gas phase and the concentration of individual reactive species. However, as the gas pressure is increased, reactions between heavy particles play a more and more dominant role in plasma dynamics. Important heavy particle reactions in plasma systems include ion–ion neutralization, ion-molecule collisions, charge exchange, and neutral reactions. Some examples of heavy particle reactions are shown below:

$$Cl^- + O^+ \rightarrow Cl + O \tag{21.7}$$

$$O^+ + O_2 \rightarrow O + O_2^+ \tag{21.8}$$

$$Ar^* + O_2 \rightarrow Ar + O + O^* \tag{21.9}$$

21.2.2.4 Surface Chemistry

Although material removal is the primary goal of plasma etching, industrial etch processes have to simultaneously satisfy many other stringent requirements. These include control of feature sidewall and bottom surface profiles, etch selectivity to other exposed materials, uniformity of etch process over large substrate surfaces, and interaction with preceding and following processing steps. These factors are discussed in detail in the next section. The complex and often conflicting requirements imposed on industrial etch processes necessitate the etch engineers to use a combination of many fundamental surface processes, some of which deposit materials instead of removing them, for etch process design. Some fundamental surface processes that contribute to etching are physical sputtering, reactive ion etching (RIE), chemical etching and polymer deposition. These phenomena are reviewed in this section.

21.2.2.4.1 Sputtering

One of the simplest material removal processes is physical sputtering.[6] Sputtering involves the bombardment of target material by energetic ions or neutrals (~50–2000 eV). Upon impact, the energetic particles transfer their energy to atoms on the target material surface. This energy is then transferred to neighboring atoms through scattering, some of which can be ejected from the surface. Sputtering processes are characterized by a threshold energy, which is typically in the range of 10–50 eV. Although sputtering is an effective technique for removing material, it tends to be non-selective and all exposed materials sputter to one degree or another. Sputtering of other than atomic targets generally changes atomic composition near the surface, as all atoms in the material will sputter at different rates. Sputtering involves energetic particle bombardment and often modifies material morphology near the surface. Sputtering of atomic targets is a well studied subject[6] and theoretical models adequately capture atomic sputtering processes.

21.2.2.4.2 Chemical Etching

Classical wet etching involves the exposure of materials to reactive fluids (e.g., acids) that react with material surface and produce volatile products, resulting in material removal or etching. Since numerous neutral radicals and atoms are generated in plasmas, there are counterparts of wet or chemical etching in

plasmas as well. A simple example of chemical plasma etching is Si etching using F,[7] which has a high etch rate even at room temperature:

$$Si_{(solid)} + 4F \rightarrow SiF_{4(gas)}$$

Another example is etching of organic polymers in oxygen plasmas, where free O atoms react with C and H in the polymer. Photoresists are one class of such polymers and are used as masks during etching. Photoresists are typically removed using O_2 plasmas after etching. Chemical etching is often isotropic as incoming neutral etchants have a uniform angular distribution. However, for some crystalline materials, chemical etching can be sensitive to crystallographic orientation and one can etch materials in preferential directions. Chemical etching tends to be very selective as it requires a unique combination of reactive gases and materials to proceed. During the fabrication of sub-micron sized features in CMOS devices, chemical etching can often not be tolerated due to its isotropic nature. Processing conditions are therefore chosen so as to minimize chemical etching. In other circumstances, e.g., micro electro-mechanical system fabrication, isotropic nature of chemical etching as well as its selectivity are essential for fabricating complex structures where etching needs to occur in all directions.

21.2.2.4.3 Reactive Ion Etching

Most plasma etching processes primarily rely on RIE for material removal. Reactive ion etching involves simultaneous bombardment of energetic ions and reactive neutral radicals on material surface. As demonstrated in the classic experiment by Winters and Coburn,[8] the etch rate is much larger in the presence of both ions and neutrals as compared to ions or neutrals alone. Since ions primarily gain their energy in the sheath above the substrate, they bombard the substrate surface almost normally. Etching therefore occurs preferentially in the normal direction (i.e., anisotropically) and so one can fabricate structures that have vertical sidewalls. Reactive ion processes tend to be less selective compared to chemical etching processes due to the presence of energetic ions but more selective relative to physical sputtering due to its partially chemical nature. The exact mechanism through which the ion and neutral radical react with the material surface is still a topic of intense research. Work in recent years has shown that, depending on the ion, neutral, and material, RIE can occur in several different manners. For a simple system such as Cl etching of Si, neutral radicals chemisorb on the material surface and form weakly bonded compounds. Energetic ions are then able to dislodge these compounds from the surface and generate volatile products. Reactive ion etching phenomenon tends to be more complex for fluorocarbon etching of silicon or dielectrics (e.g., Si_3N_4 and SiO_2).[9] In these etch systems, a thin fluorocarbon film forms on the material surface. Although etchant atoms (C and F) are present in this fluorocarbon layer, ions are necessary to modulate fluorocarbon film thickness and enhance etchant and product diffusivity through the reactive film. Surface reactions therefore do not occur without energetic ions. Reactive ion etching mechanisms are discussed in more detail in later sections in this chapter.

21.2.2.4.4 Polymer Deposition

The primary goal in most commercial etching processes is to selectively remove one material from the substrate and leave the other exposed materials intact. To generate small features, anisotropic etching also requires that etching only take place in the vertical direction with no etching in the horizontal direction. Although careful design of plasma etching reactor and appropriate choice of etching gases helps in achieving these goals, one surface mechanism that has proven indispensable in this regard is polymer deposition. Through a variety of neutral and ion assisted processes, thin polymer-like films can be deposited on material surfaces in the plasma. Presence of these films on vertical surfaces limits contact of material surface with the etchant species, inhibiting horizontal etching. Unidirectional energetic ions are however able to remove the thin polymer film on horizontal surfaces and etch in the vertical direction. Since polymer film growth rate sensitively depends on material on which it is growing, these films can be used to retard etching on certain materials and enhance etch selectivity.

21.2.2.4.5 Effect of Substrate Temperature

As is often the case with natural phenomena, many of the above mentioned fundamental surface processes take place simultaneously in etching processes. Etch engineers must judiciously choose plasma operating conditions to enhance or reduce the contribution of individual surface processes and control the final process results. One parameter that is particularly useful in this regard is the substrate temperature as many fundamental surface processes exhibit strong temperature dependence. This dependence can often be described using the Arhennius reaction expression.[10] Chemical etch rate generally increases with surface temperature. Therefore, for processes in which both physical and chemical etching components are present, one can control feature profile by varying the substrate temperature. For example, during etching of silicon using a plasma of fluorine rich gases (e.g., SF_6), it is well known that chemical etching component can be suppressed by decreasing substrate temperature and anisotropic etch profile can be obtained.[11] However, chemical component of etching starts to dominate above room temperature and the etch profile becomes more isotropic. Silicon etch rate also increases substantially in the chemical etching regime. An increase in surface temperature generally decreases sticking coefficients of neutral species on the surface and would retard polymer deposition or passivation, which can significantly modify etch selectivity. Reactive ion etching processes also exhibit substrate temperature dependence as neutral radical sticking coefficient is influenced by temperature. Substrate temperature has to be maintained at a level that none of the thin films on the substrate are damaged. One such film is the photoresist, whose glass transition temperature is normally between 100 and 250°C and photoresist starts to flow above this critical temperature. Energetic ions from the plasma bombarding the substrate during etching can substantially heat it. Wafers are therefore cooled during plasma etching using backside helium at higher gas pressure (typically a few Torr) and electrostatic chucks.

21.2.3 Plasma Etching Tools

Many plasma etch systems have been developed over the last 25 years to address the multitude of requirements posed by semiconductor etch processes. These plasma etch tools are generally based on a few basic plasma production techniques, and plasma production mechanism appears to be a suitable way to categorize plasma etchers and discuss their basic principles. The basic operating principles of capacitively coupled, inductively coupled, plasmas and, to a lesser degree, wave heated plasmas is the primary subject of this section.

21.2.3.1 Capacitively Coupled Plasmas

Some of the earlier plasma etching tools were capacitively coupled, and variants of capacitively coupled plasma etchers still dominate the semiconductor industry. In the simplest form of a capacitively coupled plasma etcher, etchant gases are injected between two metallic electrodes to which an RF voltage is applied (Figure 21.4a).[2] The potential drop across the gas breaks it down and generates the plasma. Energy is transferred from the RF source to electrons during acceleration in the inter-electrode electric field (ohmic electron heating) and collisionlessly when electrons reflect at sheath edges (stochastic heating). Significant fraction of input RF power is however consumed by ions while accelerating in the sheaths, which is dissipated at the electrode surfaces (or substrates placed on them) during ion bombardment. Electron (or plasma species) density is therefore low ($\sim 10^9$–10^{10} cm^{-3}) in capacitively coupled plasma systems, and gas dissociation fraction is also low to moderate (typically less than 0.5). Commercial capacitively coupled plasma etchers are typically operated at moderate gas pressures ($\sim$50–500 mTorr) and scattering within the gas prevents their use for fabrication of extremely small features. For typical applied RF powers ($\sim$100–3000 W), RF voltages can be a few thousand volts and physical sputtering play a major role in etching. The same RF power supply is used to generate the plasma and accelerate ions in sheaths in capacitively coupled plasma systems. The two mechanisms are therefore inter-coupled.

Crossed electric and magnetic fields have been used in many capacitively coupled plasma etchers to increase the etch rate, alleviate damage caused by energetic ion bombardment and improve etch

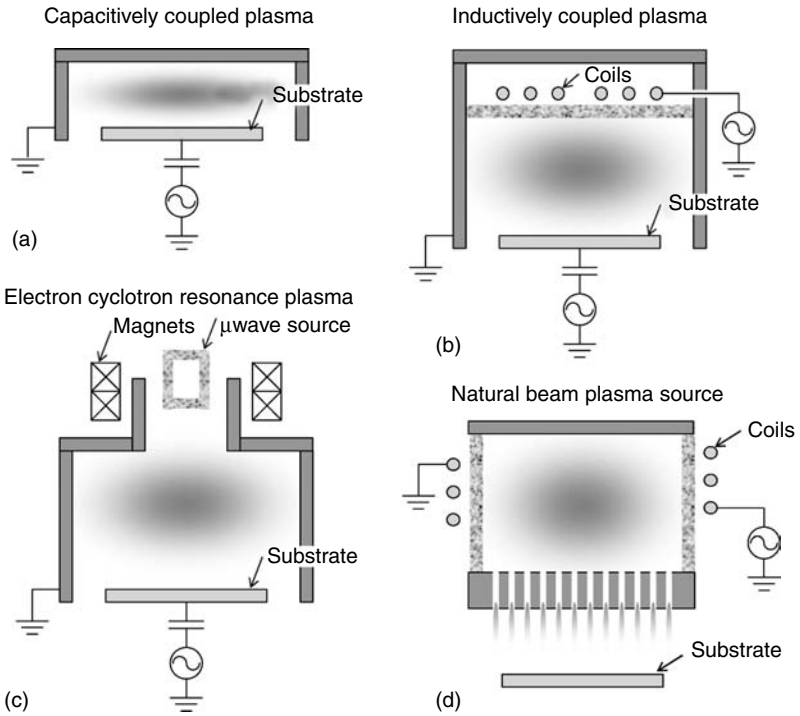


FIGURE 21.4 Plasma etching reactors: (a) capacitively coupled plasma etcher, (b) inductively coupled plasma source, (c) electron cyclotron resonance (ECR) plasma, and (d) neutral beam plasma etcher.

uniformity.[12] In these magnetically enhanced reactive ion etcher (MERIE) tools, the magnetic field reduces electron mobility towards the electrodes and their loss there. Densities of electrons and other species in the plasma are therefore larger in MERIE reactors for the same input RF power, which enhances the material etch rate. For a given RF power, higher electron (and ion) density in MERIE reactors results in lower RF voltage (or ion energy), which diminishes ion bombardment induced damage on substrate and electrode surfaces. Etch uniformity is improved in MERIE reactors by either shaping the applied magnetic field or rotating it physically or electrically. Magnetically enhanced reactive ion etcher reactors have been extensively used for dielectric etching in the semiconductor industry and they are still quite popular.

Another recent innovation in capacitively coupled plasma etcher design is the use of two (or more) RF sources at different frequencies.[13] It is well-known that capacitively coupled plasmas are generated more efficiently at higher RF frequencies and, for the same input power, RF voltages are larger at lower RF frequencies. A high frequency RF source (25 MHz and above) is therefore used in dual frequency plasma systems to generate the plasma, while a low frequency RF source (typically a few MHz or lower) accelerates ions. If the two RF frequencies are chosen appropriately (not close, no harmonic beating), the two RF sources operate in a decoupled manner over a limited parameter range. One can therefore obtain a higher plasma density relative to a simpler capacitive plasma system, and independently control ion energy as well.

21.2.3.2 Inductively Coupled Plasma

Inductively coupled plasmas (ICP) were developed in the early 1990s to address the difficult process requirements of high aspect ratio (AR) oxide etch with high selectivity.[14] Inductively coupled plasmas reactors operate at lower gas pressure (~3–50 mTorr) than capacitively coupled plasmas. Inductively

coupled plasmas reactors utilize a set of coils that are physically separated from the gas through a dielectric window (Figure 21.4b). Radio frequency current through the coils generates an electromagnetic wave that penetrates the plasma chamber, azimuthally accelerates electrons and generates the plasma. As most of the input RF power is consumed by electrons, the electron density is substantially larger ($\sim 10^{11}$–10^{12} cm^{-3}) in ICP reactors compared to capacitively coupled plasmas. In addition, a second RF source can be used in ICP reactors to separately bias the substrate undergoing etching and impart energy to bombarding ions. As separate RF sources are used for plasma generation and ion acceleration, their independent control over a broad parameter regime is permissible. As electron density is high in ICP reactors, most plasma processing gases undergo substantial dissociation (more than 90% in many cases). Although larger degree of dissociation enhances etch rate, it has detrimental effect on inter-material selectivity in many cases. For typical applied RF bias powers ($\sim$ 100–1500 W), RF and sheath voltages are up to a few hundred volts. Inductively coupled plasmas reactors are extensively used in semiconductor industry for front end etch processes (poly-silicon gate definition) and metal etching. Although ICP reactors have also been used to make significant advances in dielectric etching, they have now been replaced in many cases by dual frequency and magnetized capacitively coupled plasma etchers that have similar process capabilities.

21.2.3.3 Other Plasma Etchers

One of the most common methods for plasma generation, especially of high-temperature fusion plasmas, makes use of resonant wave-plasma interaction. Energy can efficiently be transferred from the wave to plasma in regions where resonant wave-plasma interactions occurs. Plasma medium is rich in wave phenomena, and many wave based etching-relevant plasma sources have been developed. This includes electron cyclotron resonance (ECR), helicon, and surface wave plasmas. The most mature among these for etching and materials processing applications is the ECR plasma (Figure 21.4c), whose use parallels that of an ICP etcher.[15] In an ECR reactor, microwave (typically at 2.45 GHz) is launched into a magnetized chamber containing the etchant gas at low pressure ($<$10 mTorr). Electron cyclotron resonance occurs at spatial locations where the local electron cyclotron frequency (eB/m_e) matches the applied RF frequency. By carefully designing the magnetic field profile, one can obtain high density uniform plasma above the substrate surface. Plasma densities in ECR reactors are higher or comparable to ICP reactors. Electron cyclotron resonance reactors are also operated at lower gas pressures compared to capacitively coupled plasma etchers and allow independent RF biasing of the substrate. Similar to ICP, ECR reactors are characterized by high degree of gas dissociation. Electron cyclotron resonance reactors have been primarily used for metal and poly-silicon etching.

Another class of wave heated discharges makes use of helicon or whistler waves, which are low frequency waves that can be excited in weakly magnetized plasmas.[16] Typical RF frequencies are 1–50 MHz while magnetic field varies between 20 and 200 G. Many modes of helicon waves can be excited in confined plasma systems, so a variety of antenna configurations have been developed for helicon wave source. Very high plasma densities, comparable or higher than ICP and ECR plasmas, can be obtained in helicon discharges and helicon etchers operate at low gas pressure. Helicon sources are however currently limited to research or low volume production and they have not been utilized for high volume production of integrated circuits (ICs).

DC plasma etchers were developed and utilized during early years of plasma etching and they are still viable sources for sputtering applications. The operation of a DC plasma reactor is well studied: moderate pressure gas is enclosed in between an anode and cathode, across which a DC voltage is applied. Ion acceleration across the cathode sheath leads to secondary electron emission, which sustains the discharge. The physical structure of a DC plasma can however be quite complex with many different regions having different electrical properties.[2] Radio frequency discharges have supplanted DC plasmas in recent years due to higher ion energy, higher gas pressure and uniformity related issues in DC discharges.

Because of the presence of charged species or ultraviolet radiation in plasma, electrical damage of circuits on the substrate remains a constant concern during plasma etching. To alleviate this problem,

plasma etch sources that rely on energetic neutral beams have been developed in recent years.[17] A typical neutral beam source is shown in Figure 21.4d, and similar designs have been used in the past for ion milling applications. Plasma of the appropriate gas is generated through conventional means, which is then allowed to seep through holes in the electrodes. Ions can then be accelerated and neutralized before they bombard the substrate, resulting in energetic neutral species bombardment on the substrate. Neutral beam sources are under development currently and they have not been used for high-volume production. However, as requirements for higher anisotropy and selectivity increase, neutral beam sources may be able to offer new solutions.

There are numerous etching applications where etch processes can be adequately performed by reactive neutral species that are generated in the plasma. For such applications, another technique that has been utilized to alleviate plasma charge damage is to generate the plasma remotely, i.e., away from the substrate, and transport neutral species to the substrate so that ions are either excluded or neutralized. Remote plasma sources or chemical down-stream etchers are used for many plasma cleaning and material treatment applications. These etchers are also useful for high etch rate removal of blanket films that do not have any patterned features. Their application for anisotropic etching applications is however limited due to the broad angular distribution of neutral etchants.

21.2.4 Plasma Etch Issues

Integrated circuit manufacturing is truly a batch processing operation where billions of sub-micron sized devices are fabricated and connected simultaneously on large substrates (up to 300 mm diameter currently). The need to reliably manufacture all these devices imposes a set of stringent requirements that all IC fabrication processes, including etching, need to satisfy before they can be used for manufacturing. A detailed understanding of these requirements as well as the tradeoffs involved is essential for etch process engineering. In this section, we describe some factors that are considered important for an etch process' manufacturability, and need to be paid attention to when an etch process is designed.

21.2.4.1 Uniformity

A large number of ICs are produced simultaneously on large wafers (up to 300 mm diameter) and manufacturing operations need to treat all ICs identically irrespective of their location on the substrate. All manufacturing operations, including etching, therefore need to process the substrate uniformly over its whole area. This includes etch rate of all exposed materials and feature profiles for structures of varying dimensions. Although plasma operating parameters (e.g., RF power, gas pressure, gas flow rates) do impact the plasma structure, close to perfect uniformity can only be achieved if the plasma etching reactor has been carefully designed. In many plasma etch systems, gases undergo substantial dissociation and gas residence time determines how much and where do the feed gases fragment. Plasma reactors therefore need to be designed so that gas inlet and pump ports are symmetric, and lead to spatially uniform dissociation or ionization of the gases. Low gas pressure operation (e.g., in ICP and ECR reactors) and plasma production away from the substrate helps improve uniformity by allowing species density to homogenize through diffusion. In capacitively coupled plasma reactors, one large source of plasma non-uniformity is electric field enhancement at electrode edges. This field enhancement leads to non-uniform power deposition above the substrate, leading to higher plasma densities and etch rates near substrate edges. Non-uniformity due to field enhancement can be mitigated either by making the electrode larger than the substrate or adding materials at electrode edges (e.g., focus or shadow rings) that reshape the electric field profile there. Magnetic fields in MERIE reactors can introduce non-uniformities in the plasma through $E \times B$ drifts. These non-uniformities have been minimized by either (electrically or mechanically) rotating the magnetic field or designing magnetic field in a manner that $E \times B$ drift is minimized. One large source of non-uniformity in ICP reactors is localized plasma production, which is a function of current in the antenna coils. Careful attention has therefore been paid to the design of antennas and their RF driving circuitry. Electrode edge electric field enhancement can cause uniformity problems in ICP reactors as well, and strategies for alleviating them are similar to those

used in capacitively coupled plasma etchers. Even in a well-designed plasma reactor, plasma etch uniformity will vary as a function of operating conditions. For example, decrease in gas pressure typically improves uniformity by allowing more homogeneous gas distribution. Identification of process window that provides adequate etch uniformity is an essential part of etch process design.

Highly electronegative plasmas popular for silicon etch are prone to a high degree of non-uniformity given propensity of the negative ions to pool in the center of high density plasma sources. Diluents that can change the plasma from electronegative to more electropositive are a means of achieving greater uniformity. Examples of additives are helium and argon. The trade-off is that the more electropositive plasma with massive ions like argon could lead to increased sputtering that in turn leads to corner faceting and more micro-trenching. Helium is a reasonable option given its lower sputter rate.

21.2.4.2 Selectivity

The primary purpose of most etch processes is to etch a particular material and leave the other exposed materials intact. The materials that should ideally not be etched include the mask, which should retain its integrity during the etch process, and underlying materials exposed during the etch process (e.g., material B in Figure 21.1). A key parameter defining the efficacy of an etch process is its selectivity to other exposed material, where selectivity is defined as the ratio of the etch rate of desired material vs the etch rate of the other material. Chemical etch processes are characterized by a large selectivity if the reactive radicals are chosen so that they only react with the desired material. Extremely large selectivity is however difficult to achieve in RIE processes as energetic ions can sputter or etch all materials to one degree or another. Many techniques are utilized to distinguish materials during etching and obtain acceptable selectivity. One such methodology is to deposit polymer or protective layer on materials that do not need to be etched. For example, when a stack of SiO_2 on silicon is etched at moderate ion energy (e.g., in an ICP etcher) in a fluorocarbon plasma, SiO_2 is a readily reactive ion, etched through the synergistic effect of fluorocarbon radicals and ions. However, when underlying silicon is exposed, fluorocarbon radicals react on silicon surface and generate a relatively thick polymer film there. This polymer film protects silicon from further etching. Following the etch process, the fluorocarbon protective film can be removed in an O_2 plasma along with the photoresist. All ion assisted etch processes are characterized by a threshold, where no etching takes place below the threshold energy. Another method for enhancing selectivity utilizes the relative difference in material etch thresholds. By choosing ion energy in between these etch thresholds, one can etch the desired material at a reasonable rate while leaving the other material virtually intact. Selectivity can also be enhanced by choosing etchant gas(es) so that radicals only react with the desired material but are non-reactive with the other material. The etch process can be quite selective since the etch process for the second material layer will be primarily physical sputtering and this typically has a much lower etch rate than RIE. This is the case with etching of silicon on SiO_2 in Cl_2 or HBr plasma. Silicon readily etches in these plasmas but SiO_2 has a relatively low etch rate.

21.2.4.3 Loading and Isolated-Dense Bias

Most ICs have complex patterns on them, and feature density will vary appreciably from region to region. If the etch process has been designed to maximize the etch rate, it is expected that regions with high concentration of etchable structures or exposed area will consume more of the etchant species. If reactive etchants cannot be replenished in the plasma at a fast enough pace, etchant density will decrease in the vicinity of dense features, decreasing the local etch rate. On the other hand, isolated structures have an ample supply of etchants and so will etch at a relatively higher rate. This disparity between the etch rate of isolated and dense features, commonly known as iso-dense bias, can negatively impact the etch characteristics. This is especially true if selectivity to underlying materials is not high. Since etch time is chosen to completely clear primary material in regions with the lower etch rate (typically dense), underlying material will get exposed in regions with higher etch rate (typically isolated structures) towards the end of the etch process and may get damaged during the long exposure to the plasma. The above picture of iso-dense bias due to etchant depletion can get complicated if polymer deposition plays a role in the etch process. Iso-dense bias can typically be reduced by moving to a process regime where the

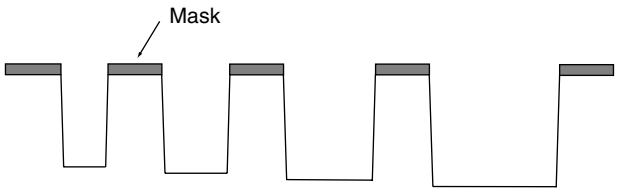


FIGURE 21.5 Impact of feature size on etch rate.

supply of etchants is substantially more than their consumption, e.g., by increasing flow rate of etchant gases.

21.2.4.4 Aspect Ratio Dependent Etching

In addition to feature density variation, structures of widely varying dimensions can be present on the substrate during etching. Since the etch rate often depends on feature size, problems similar to iso-dense bias can occur. A classical example of feature size dependence of etch rate is aspect ratio dependent etching (ARDE), which is illustrated in Figure 21.5. Consider RIE of an array of vias due to the synergistic effect of neutral radicals and energetic ions. Ions have a narrow angular distribution at the substrate surface due to acceleration in the sheath, and their flux at the bottom of vias will not depend on opening size. On the other hand, neutral radicals have a broad angular distribution at the substrate surface, and they will be partially blocked by via sidewalls from reaching the bottom surface. As a consequence, narrower structures will etch more slowly compared to wider openings and, over time, the bottom surface will become non-uniform. The ARDE picture of course becomes more complicated if polymer deposition occurs simultaneously with etching. Since ARDE is very much a geometrical issue, plasma parameters typically may not have a strong impact if RIE is the dominant surface process. However, modification of process conditions such as using a more polymerizing chemistry can be used to reduce the ARDE effects.

21.2.4.5 Feature Profile and Sidewall Shape

Although some applications require different post-etch profiles, Figure 21.6a shows a profile that would be considered "ideal" for many applications. In this profile, the sidewalls are straight after etching and the bottom surface is flat. Since most etching processes are a combination of RIE, chemical etching, and polymerization/passivation, this ideal profile is often unachievable. Understanding profile imperfections that can occur during etching and how they can be corrected is invaluable for etch process design. Source

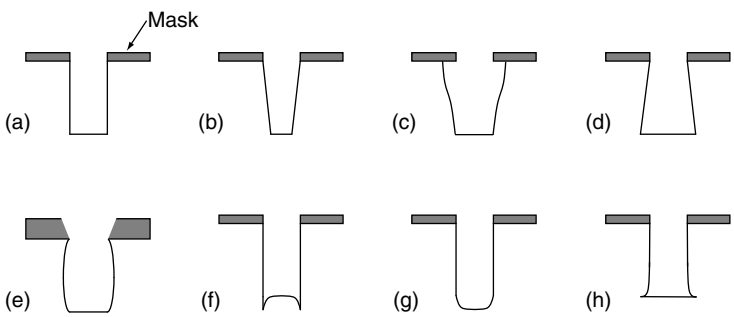


FIGURE 21.6 Feature profiles in an etch process: (a) vertical sidewalls and bottom surface, (b) tapered sidewalls, (c) under-cutting, (d) reentrant profile, (e) bowing, (f) micro-trenching, (g) curved bottom surface, and (h) notching.

of profile imperfections are typically two: artifacts of individual surface phenomena, and distortions generated by reflection of ions and neutral species off the feature surfaces. Some commonly observed feature profiles are shown in Figure 21.6. If plasma promotes polymer deposition on feature sidewalls, polymer deposition reduces etching at corners which tends to gradually narrow the opening. This leads to tapered sidewalls (Figure 21.6b). Integration issues sometime make slightly tapered sidewalls preferable over vertical sidewalls since conformal deposition of subsequent films is easier. Excessive polymerization can however, lead to significantly tapered sidewalls and in extreme cases, etching can stop. On the other extreme, undercutting can occur (Figure 21.6c) if chemical etching occurs on lateral sidewalls and there is inadequate sidewall passivation. Undercutting increases the width of feature opening and shrinks the material in between openings. Exposed corners can become tapered in sputter dominated processes whose yield typically peaks at off-normal angles. If sidewall protective films are being deposited by "sticky" neutral species, the flux of these species will decrease with height in narrow, large aspect ratio features. There will therefore be little protection near the bottom of the sidewalls, resulting in a reentrant profile (Figure 21.6d). Reentrant cavities are one of the most undesirable post-etch profiles from the integration perspective as it becomes difficult to deposit conformal material in them without voids. Reflection of ions and neutrals from the feature surfaces can also distort the feature profile. One such anomaly, which is illustrated in Figure 21.6e is called bowing and it occurs if mask sidewall is tapered and ions can reflect off them. Bowing can also be caused by multiple bounces within features of reactive but less than unity sticking coefficient reactive radicals such as oxygen, which lead to erosion some depth below the feature entrance. Energetic ions bombard the substrate almost vertically due to acceleration in the sheath. If sidewalls have slight taper, these energetic ions impact the sidewalls at an almost grazing angle and can specularly reflect off them. Resultant enhancement in ion flux near feature edges enhances the etch rate there leading to creation of micro-trenches (Figure 21.6f). Local non-uniformity in protective polymer flux at the bottom of features, e.g., due to shadowing or ARDE effects, can also cause micro-trenching. If a RIE process is operated in the neutral starved regime, reduced flux of neutral etchants at feature corners can reduce the etch rate there leading to rounding of the bottom surface or the formation of footers (Figure 21.6g). Charging of the feature surfaces and mask etching can further distort the feature profile and these issues are discussed in the following sub-sections.

21.2.4.6 Mask Interactions

From the etching perspective, it would be ideal to use a mask that is thin, is not impacted by the etch process, and is easily removable following etch. Unfortunately, photoresist masks often do not satisfy any of these etch requirements. Interaction of the mask with film being etched is therefore important to consider in etch process design. As the mask will erode during the etch process, masks need to be thick enough that they do not lose their integrity during etching. An excessively thick mask will however make the effective aspect ratio unnecessarily large, increasing the propensity for etch stop or feature sidewall distortion. High aspect ratio is also difficult to achieve during the photoresist development process. Attempts are therefore being made to determine the minimum mask thickness necessary for each etch application. Like many other materials, sputter yield of photoresists is largest at an off-axis angle ($\sim 45°$). This results in tapering of photoresist mask corners during etching. Ion and neutral radical reflection off these tapered edges can distort profile of the etched structure. If sputtering is accompanied by polymer deposition on photoresist sidewalls, several facets at different angles can form, further distorting feature sidewalls. The worst impact of mask faceting is felt if the facet approaches the underlying material, resulting in substantial increase in feature opening at the top. Photoresists have conventionally been removed or ashed using O_2 plasmas. Although O_2 plasma does not adverse impact traditional dielectric materials (SiO_2, Si_3N_4), there is ample evidence that O_2 plasmas can damage the surface of low-κ dielectrics increasing the effective dielectric constant. Taking into account, the feature distortion caused by the mask and post-etch processes, has therefore become essential for etch process design. As the industry migrates towards lower wavelength UV radiation [157 nm, extreme ultraviolet (EUV)], the thickness of photoresist that can be adequately imaged is rapidly shrinking and etch resistance of these materials is significantly reduced as well. The reduced photoresist budget has led to the development of

many complex mask strategies. In double and triple layer schemes, top layers are thin but can be photolithographically patterned. Following lithography, plasma etch is used to transfer the mask pattern on top thin layer to the thicker under-layers. These under-layers then act as the mask during the primary etch process. Double and triple layer masking strategies have been possible due to the development of new polymeric materials, and will probably be required for nanoscale feature patterning unless totally new materials are developed.

21.2.4.7 Charging

By their very nature, plasmas contain charged species such as positive ions and electrons. These charged species come in contact with the substrate and, if the material stack on the substrate does not allow ready neutralization of charge, large electric fields can build up. These electric fields can damage the structures on the IC in numerous ways. First, energetic ions play a crucial role in most RIE processes. Electric fields due to feature charging can distort ion trajectories within the feature, making them bombard on surfaces, which are not intended to be etched. Resulting etch profiles are therefore substantially different from those desired. A classic example of this charge damage phenomena occurs during the overetch step of poly-silicon gate etch.[18] When poly-silicon is completely removed during gate formation, the underlying SiO_2 dielectric film does not allow the unequal charges to neutralize on the wafer surface. The resulting ion-trajectory-distorting electric fields lead to notching at the bottom of the gate structure (Figure 21.6h). Currents from the plasma can also generate excessive electric fields across thin gate dielectrics, stressing them beyond the point of recovery. Resulting damage often manifests itself through reduced process yields and degraded device reliability. If charging is suspected to be the cause of problems during etching, etch process can often be moved to other regimes where charging is less prevalent. Integrated circuit layout can also be modified to reduce antenna ratios (the ratio of current collection area to gate dielectric area).

21.2.4.8 Integration Issues

Materials etching is one component of the complete process of fabricating ICs, and what happens during etch is invariably linked to preceding and subsequent steps. As mentioned earlier, slight taper in the sidewall profiles is desirable in many cases as it helps during subsequent deposition steps (e.g., using physical vapor deposition (PVD) or its ionized variants). Opening of trenches or vias might be even further broadened to assist the subsequent deposition processes. Etching of very small features in dielectrics has increasingly become more complex and the process might modify material at the feature surface or leave contaminants there. Therefore, before subsequent steps, surface often needs to be treated or cleaned. Surface pores also need to be sealed for porous dielectric materials.

21.2.4.9 Environmental Health and Safety

As massive amount of consumables are used during IC manufacturing, environmental health and safety issues have become an important consideration for process design. As these issues are discussed in detail elsewhere in the handbook, we only briefly comment here on some etch relevant issues. Etching utilizes gases that are often toxic, corrosive, or contribute to global warming. Many of these gases survive for a long time in the atmosphere (e.g., Half-life (CF_4) = 50,000 year, Half-life (c-C_4F_8) = 3200 year) in the atmosphere and have high global warming potential (e.g., GWP (C_2F_6) = 12500, GWP (SF_6) = 24,900). It is obvious that we cannot afford to release un-reacted etch gases in the atmosphere. Effort is therefore underway to reduce harmful etch relevant environmental impact through abatement and process optimization. Abatement devices are being developed that treat etch effluents and transform them into less harmful substances.[19] These abatement devices can however be expensive and add their own environmental risks by consumption of combustible materials and use of large quantities of water. In addition, etching gases with lower GWP are being developed (e.g., C_5F_8, C_4F_6) and are replacing the more environmentally harmful gases. In some cases, these alternate chemistries have shown to offer process advantages in addition to lower GWP.

21.3 65–90 nm CMOS Etch Process Modules

In CMOS IC manufacturing, plasma etching is usually dealt with in the context of a process module or grouped steps of manufacturing operations to form a functional structure on the wafer. The fabrication of the transistor, commonly referred to as FEOL manufacturing, comprises of several modules including gate etch and shallow trench isolation (STI) etch. The wiring from the transistor to the package in the BEOL comprises of trench and via etches. In Section 21.2, the importance of plasma etching in successful BEOL and FEOL integrations was emphasized. In this section, the engineering issues associated with the key CMOS, BEOL, and FEOL modules are used to frame state-of-the-art plasma processing methods.

We first focus on FEOL etch processes, which include STI etch, gate fabrication, and contact etch. Done prior to gate fabrication, STI is the means by which active areas are electrically isolated from one another. The isolation is brought about by depositing an insulating layer in a shallow trench with the goal of retaining overall planarity. Shallow trench isolation etching involves two critical steps: the patterning of the defining (e.g., nitride) hard mask and etch of the underlying silicon. As contact films (e.g., silicides) and other materials are deposited over the dielectric, planarity may be crucial as etch processes designed to remove these films may not clear them where there is topography.[20] Issues potentially related to STI etch are stress generation, leakage, and film encroachment (with stress generation) when isolating films are grown in the trench. An illustrative STI flow is included in Figure 21.7.

A process flow for gate stack formation is described in general in Figure 21.8. The process starts with the patterning of silicon deposited over a layer of dielectric (conventionally thermal oxide) on a silicon wafer or a wafer with buried oxide (BOX). Patterning "fine" dimensions is now usually done by patterning with photoresist on top of an antireflective coating (ARC) layer. A hard mask between the ARC and the photoresist is sometimes necessary as the etching properties of the ARC layers critical for photolithography may be similar enough to the photoresist. Advanced integrations may see the ARC behave like a hard mask itself in subsequent etch steps. Antireflective coatings may be organic or inorganic with their ultimate removal process depending on their nature and their selectivity to the material above (e.g., photoresist) and the material below. Use of ARC layer as mask lowers the aspect ratio and lag associated with the following etch processes. If the ARC layer is used as mask, ARC etch step is followed by a breakthrough step that makes the ARC sidewalls more vertical, clearing what are essentially the same as footers to be discussed when overetch is discussed post the main gate patterning etch.

The gate itself must be etched in several steps due to the need to achieve a specific gate profile and clear to but not damage the underlying dielectric. These steps are known as the main etch, soft landing and overetch. In the main etch, anisotropy is usually achievable with plasma chemistries that sufficiently polymerize or protect the sidewalls although in chlorine chemistries anisotropy can be obtained taking advantage of the directionality of incident ions only. In the discussion that follows, references to literature process studies will show how this anisotropy is achieved.

In general, feature proximity and aspect ratio effects can cause variation in etch rate within different features and profile variations even as severe as notching. Minimizing these variations is achieved through masking strategies (e.g., through the use of hard masks) and process alterations. Generally, to achieve a high etch rate with anisotropy, a relatively large bias power is applied to the wafer during the main polysilicon etch step. This large bias can damage the underlying substrate so it is usually followed by a soft landing step that opens to the underlying dielectric. Footers are often the final result of the main etch step. These footers can result from the sidewall scattering processes and shadowing phenomena described in Section 21.2.4.5. An overetch step is used to knock back the footers. Proximity effects leading to lag or simply differences in the sidewall etch characteristics can lead to re-entrant profiles. Proximity effects combined with non-uniformity and the extra time required to clear material on certain regions of the die or wafer may lead to the lateral etch of the bottom of the gate creating notches. The accumulation of charge due to electron shadowing may itself result in notching as well.

To make electrical contact to the gate, source, and drain with low enough contact resistance, contacts made of silicides are fabricated to connect the metal (e.g., tungsten) and the gate, source, and drain

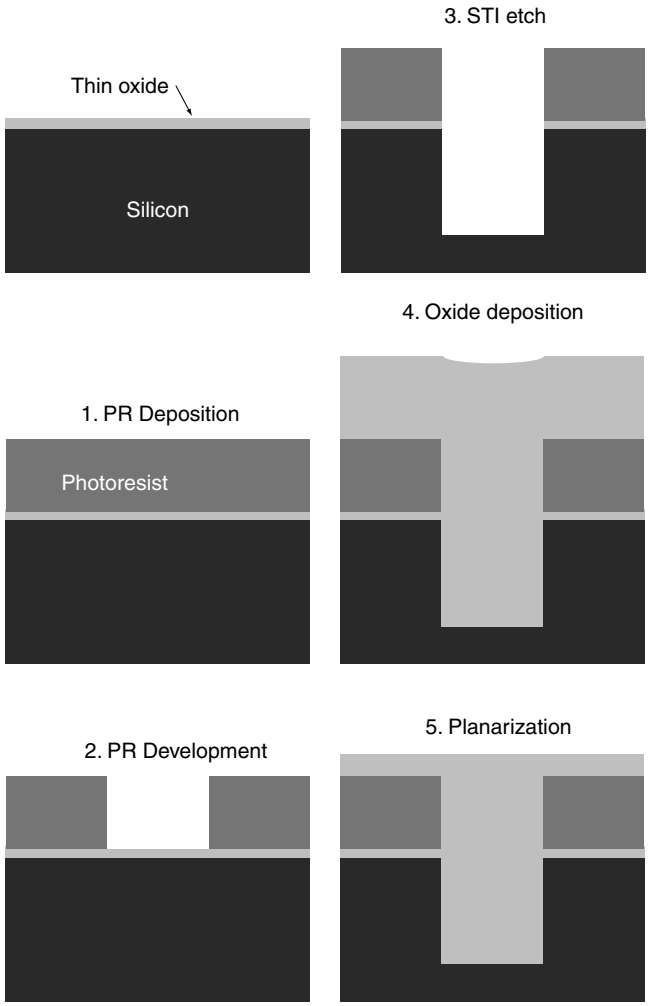


FIGURE 21.7 Steps in a typical shallow trench isolation (STI) etch process.

regions. These processes may (or may not) themselves employ dry etch processing. Connecting the contacts to upper level metals involves contact etch [usually a self-aligned contact (SAC)] through a dielectric material (potentially of high aspect ratio), which must be highly selective to a conformal etch stop layer (for example, silicon nitride) protecting the contact and the spacer. In addition to the selectivity requirements step, structures can result deep in the contact and retaining the integrity of the contact layer while making a clean landing after removal of the etch stop layer is paramount. Conditions that may promote a clean surface for subsequent metallization may be at odds with processes designed to retain the presence of thin contact layers. A basic SAC etch flow is described in Figure 21.9.

21.3.1 Shallow Trench Isolation Etch

In addition to deep well implant isolation, electrical isolation between transistors has evolved from local oxidation of silicon to BOX Isolation and STI, which involves the etching of trenches and filling them with an electrically isolating dielectric material. STI involves blanket deposition of a nitride layer that acts as the mask during Si etching, etching the masked nitride layer, etching silicon, deposition

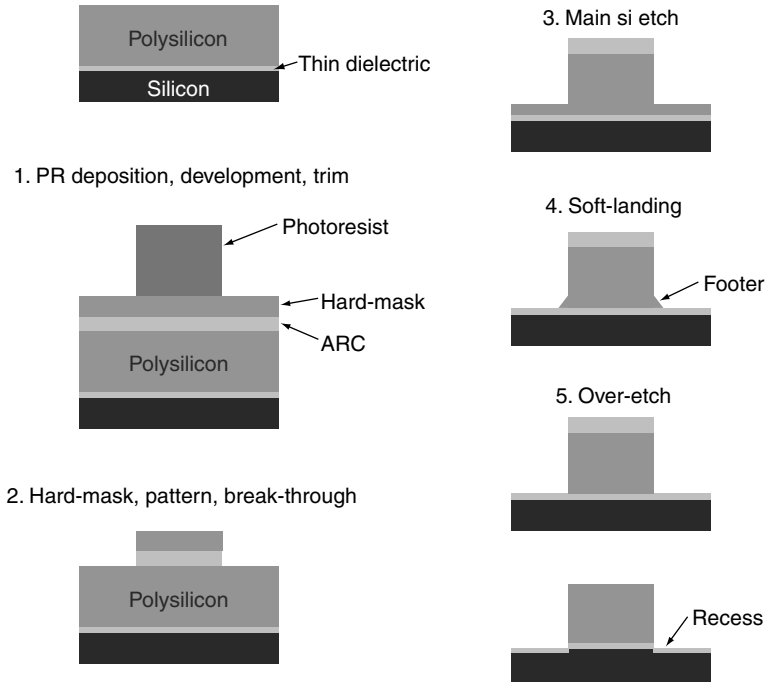


FIGURE 21.8 Steps in a typical polysilicon gate etch process.

[chemical vapor deposition (CVD) or plasma enhanced chemical vapor deposition (PECVD)] of the dielectric, and planarization. Variations in integration typically involve how many chemical mechanical planarization (CMP) steps are required in the above to initiate planar surfaces for deposition and etch.

The sidewalls resulting from the nitride etch should be as vertical as possible and the process should be reasonably CD insensitive to prevent CD shift. Nitride etches that meet this requirement can be CF_4/CHF_3-based, for example. Requirements for the Si etch typically are that the trenches have slightly sloped sidewalls, rounded corners at the top and oxidized bottom surface. The reason for the sloped wall requirement is that CVD and PECVD fill processes may evolve a seam. Among other consequences, the seamed CVD dielectrics will etch and planarize differently than unseamed trenches leading to aspect ratio dependent effects. Sharp corners correlate with increased levels of electrical leakage.[21] In order to obtain off-vertical sidewalls and rounded corners, it is important that the nitride layer capping the silicon not be etched into the silicon as it will initiate vertical sidewalls. A further requirement of the silicon etch is that the selectivity to nitride be uniform across the wafer so that CMP margin is preserved.

Obtaining rounded corners and off-vertical sidewalls is a consequence of material build-up on the sidewalls, as etching evolves. This relationship between material build-up on sidewalls and sidewall profile is true for all anisotropic etch processes. An example STI silicon etch system may include Cl_2–HBr–He–O_2, as described by Yeon and You,[22] who observe that SiO_x films (probably containing Cl and Br as well) are deposited on the sidewall as etching proceeds. In their experiments, developed photoresist is used to pattern a silicon nitride film which behaves as a hard mask. The process is therefore intended to be selective to the nitride hard mask layer. More He–O_2 in their recipe leads to more taper and higher selectivity to nitride.[22] The higher selectivity they observe is probably related to deposition of etch products combined with oxygen and other etch gas species on the nitride. A schematic representation of film buildup and the impact on STI profile is shown in Figure 21.10 illustrating the relationship between polymer buildup and taper. The deposition of material on sidewalls is less favored with increasing aspect

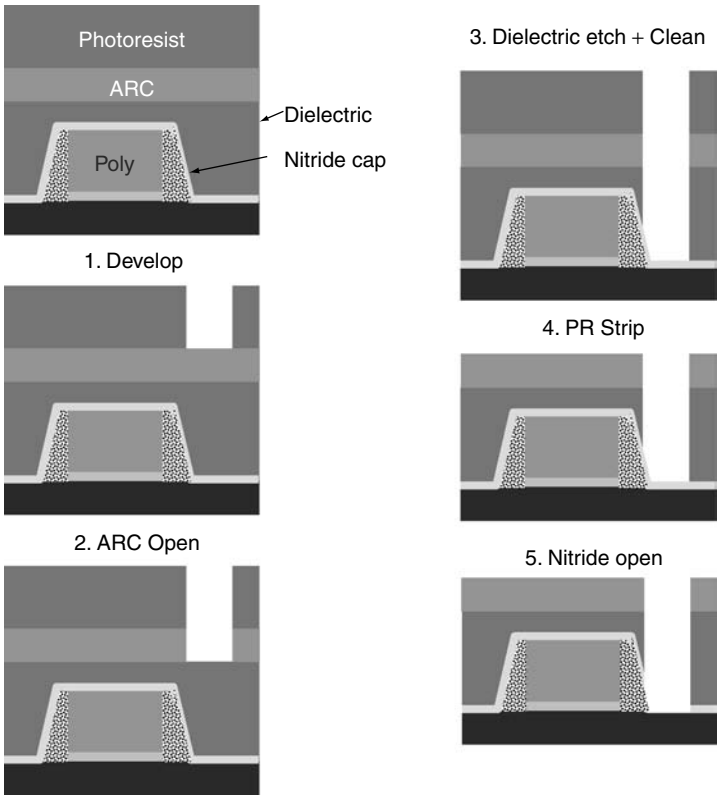


FIGURE 21.9 Steps in a typical contact etch process.

ratio or closer feature-to-spacing due to shadowing of sidewall build-up precursors from the plasma that have their origin possibly on chamber walls or the horizontal surface of the nitride itself. Hence sidewalls tend to be more vertical for smaller CDs. Taper is plotted as a function of spacing in Figure 21.11. As a quantitative example of sidewall film growth relative to etch rate, Ullal et al.[23] show that for silicon etch rates on the order of 2000–4000 Å/min, $SiCl_xO_y$ deposition rates on the sidewall are on the order of 100–200 Å/min. [Lam Research Corporation (LAM) TCP, ~400 W source power, ~50 W wafer bias power, 10 s mTorr, ~100 sccm/5 sccm Cl_2/O_2].

21.3.2 Gate Stack Etch

The objective of the gate stack etch process is the construction of the transistor gate structure by etching polysilicon selective to an underlying gate dielectric layer. Patterning the polysilicon is enabled either through the use of a photoresist developed on an ARC or a hard mask which overlies the ARC. The photoresist CD may be reduced below what is possible using optical lithography alone using a plasma etch process called trimming. The ARC may be etched as part of the trim process or etched in a separate step.[24] The ARC may function as a hard mask at some point during or after the polysilicon main etch step.

Polysilicon is usually etched with halogen (Cl, Br, F) based plasmas. With respect to the etching of polysilicon selective to underlying dielectric, chlorine is chosen to provide a reasonably high etch rate with anisotropic profile resulting from the directed nature of the ions. By itself, Cl_2 has drawbacks such as a propensity to lead to microtrenching. O_2 and HBr are added to aid in passivating the sidewall to promote anisotropy (and added selectivity to the underlying oxide).[25] Fluorine

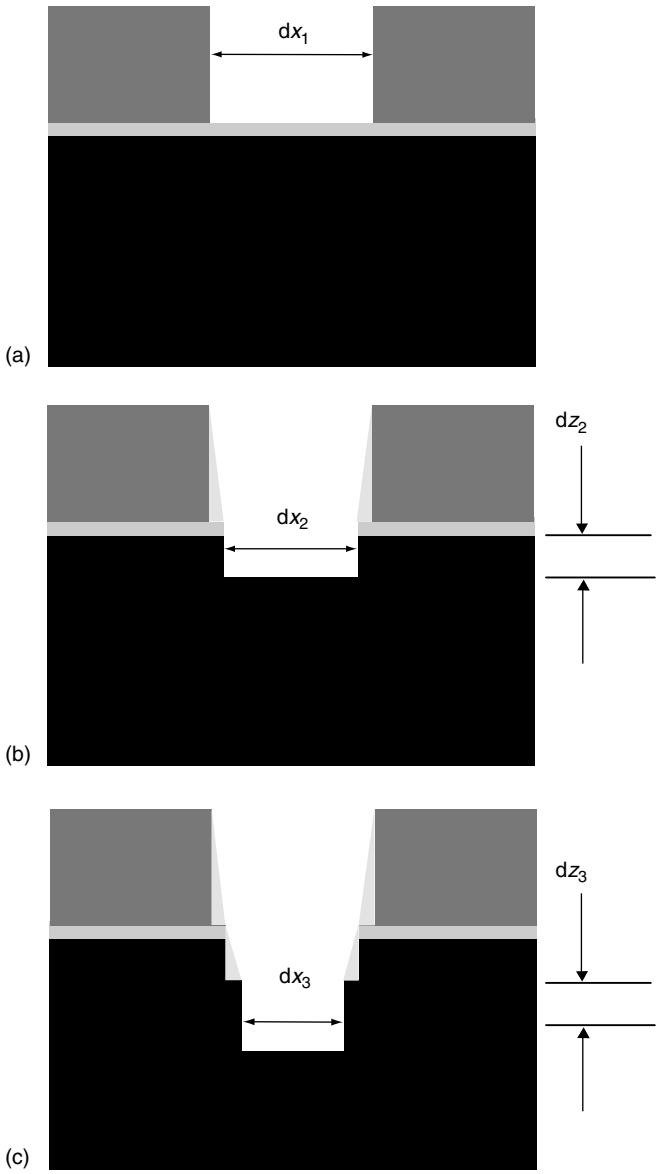


FIGURE 21.10 Idealized shallow trench isolation taper mechanism. An initial opening dx_1 is narrowed to dx_2 after etching an amount dz_2. The polymer resulting from etching to a depth of dz_2 behaves as a mask for the etch of an additional depth dz_3 resulting in a narrower opening dx_3. (From Yeon, C.-K. and You, H.-J., *J. Vac. Sci. Technol. A*, 16, 1502, 1998.)

spontaneously etches silicon and F-based plasma chemistries, by themselves, promote isotropic etching of polysilicon. Fluorocarbon additives like CF_4 are added to other fluorine sources or are used as the primary fluorine source to provide carbon based polymer for sidewall protection. Fluorocarbons may also be added to chlorine plasmas for sidewall integrity preservation. SF_6, an ample source of fluorine radicals, has been used with additives like those just mentioned for gate profiling. Complicating matters are the difference in etch rates for *n*-type, *p*-type, and undoped

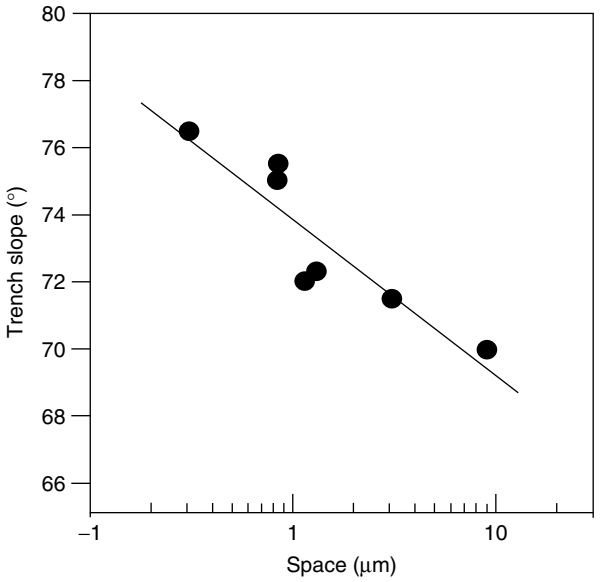


FIGURE 21.11 Dependence of ultimate sidewall slope on pattern density, where 90° is vertical. Open trenches are characterized by more taper. (Reprinted from Yeon, C.-K. and You, H.-J., *J. Vac. Sci. Technol. A*, 16, 1502, 1998. With permission. Copyright 1998, American Institute of Physics.)

polysilicon and the relationship between etch processes and gate dielectric electrical integrity mediated through effects like sidewall roughness.

21.3.2.1 Photoresist or Hard Mask Trim Etch and Gate Dimension Thinning

Photoresist etching is a complex topic given the plethora of single, bi-layer, and tri-layer photoresist schemes and the advent of 193 nm (ArF) resists and immersion lithography. Other than the work of Greer et al.[26] regarding photoresist fluorination, phenomenological, and fundamental models of photoresist etching are largely absent. Some basic observations are useful. One hundred and ninty-three nanometer photoresist is known to etch faster than the 248 nm photoresist. One hundred and ninty-three nanometer photoresist is not aromatic whereas the 248 nm photoresist is, which decreases the volatility of ArF photoresist in a plasma. Higher densities of carbon–oxygen bonds and carbonyl groups are also more volatile.[27]

In trim etch, the goal is to provide if possible, a faster lateral etch compared to the vertical etch rate to reduce the effective CD while maintaining as much photoresist thickness as possible for subsequent patterning etch steps. This is achieved either by etching in a plasma that would potentially cap or deposit on top of the photoresist while still etching laterally, or by the use of a neutral driven etch process such in an ICP source with very low bias to control sidewall roughening. A typical trim etch process includes O_2, potentially with CF_4, HBr and Cl_2 for process optimization purposes.[28] An example of a trim etch is shown in Figure 21.12. In the figure a photoresist stump is "trimmed" for different time periods. Finally, in Figure 21.12, the stump is eroded due to faceting at the corners and its structure is compromised.

How the trim process integrates with the ARC or hard mask and their individual etch processes is important to the final gate structure profile. Antireflective coatings under the photoresist significantly reduce the reflected light through absorption or destructive interference during photolithography, thus improving process resolution and contrast. Antireflective coatings can be either organic polymeric material similar to the photoresist but formulated to be more absorbing at the required wavelength or can be an inorganic layer such as an oxide or nitride. Common classes include dielectric ARCs like

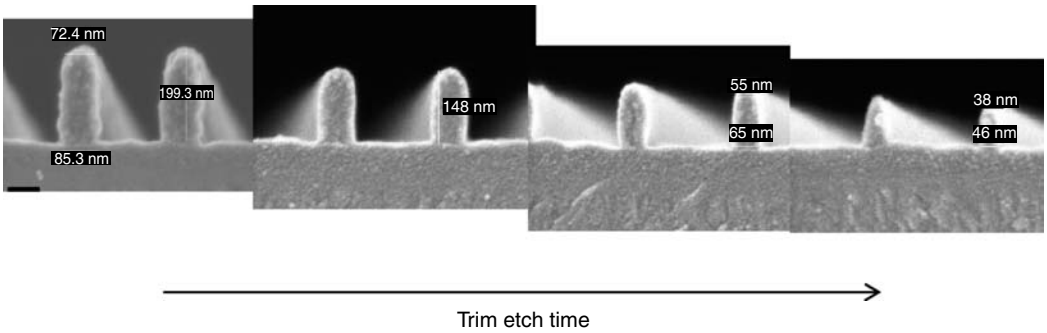


FIGURE 21.12 Evolution of photoresist line during trim etch. The left most photoresist stump represents a starting condition. Moving from left to right, the photoresist experiences more trim time.

SiON,[29] carbon containing ARCs,[30] or silicon containing organic ARCs, typically polysilanes.[31] In complex arrangements, two ARCs can be used, one as a true anti-reflection medium and another as an absorbing medium.[32] The use of ARC layers can significantly improve the lithography process performance but at a significant impact on the etch process. Organic ARCs are removed using the same chemistry that removes the photoresist so there is always a difficult tradeoff to be able to remove the ARC layer without impacting the quality of the lithography to etch the underlayer. Inorganic ARC layers are etched with highly selective chemistries so the lithography is not impacted as much. However, if the inorganic ARC is thick, the overall feature aspect ratio increases, making etching more difficult. Antireflective coating etch must also not have a deleterious impact on line edge roughness (LER). Depending on the specific ARC, another material is placed below the photoresist either for the protection of the ARC or to protect the photoresist from chemical interactions with the ARC. Especially if the ARC itself plays the role of a mask during polysilicon etch, ARC must be cleared without a footer remnant.

Templating roughness from trimmed photoresist by the ARC or hard mask layers is important for roughness in the final gate profile. Process related roughness dependencies have been characterized by Rauf et al.[33] and Goldfarb et al.[34] The use of multilayer patterning schemes can also help reduce LER by providing more independent control and modification of each step of the process. Since a critical factor for almost any resist material is LER increase with exposure a multilayer scheme with a thin imaging resist layer can be resolved with a very short exposure. A very thin transfer layer such as oxide can than be etched without further increase in the LER. Since transfer layer can be used to have very high etch selectivity to the underlayer etch mask material, such as an organic or carbon layer, the resultant LER is less since the intermediate mask layer does not erode as rapidly in the etch than the resist material would in the longer etch step. As seen with silicon containing bilayer resist, the dry develop can be optimized to significantly reduce LER somewhat independent of the exposure. Post-treatment of photoresist have also been used to make chemically amplified photoresist more resistant to the etch chemistries, by effectively repairing the polymer structure.[35] In addition, LER must be considered as primary requirement for the optimization of each etch processes beyond development and trim.

The chemical nature of individual ARCs determines how they are etched. Organic ARCs are etched with O_2 based plasmas. Etching of Si containing ARCs, whether organic or inorganic, utilizes some halogen typically. The ARCs, again, must integrate both with the photoresist etch process and the etch process of the material below it, possibly for several etch steps depending on the role the ARC plays.

Xu et al.[36] discuss the coupled roles of chamber condition and plasma chemistry on ARC etch process control. They compared the etch behavior of an ARC on the chamber seasoning condition (clean anodized aluminum or oxide coated) and gas mix (O_2 with HBr/HCl/Cl_2). As oxygen was the primary

etchant, they found that processes that enhanced the recombination of oxygen radicals on surfaces (favored on anodized aluminum over oxides) or the production of OH on surfaces (mediated by HBr and HCl not Cl_2) would lower the etch rate. Lower ARC etch rates led to CD increased due to redeposition. The presence of H (from HBr) in the plasma can actually enhance the photoresist etch rate suggesting a complex integration relationship.

The chemical properties of more complex ARCs such as polysilanes can impact process integration as well.[31] Polysilanes are thought to degrade in chlorine plasmas (etched with Cl_2 because of the significant Si content which ranges between 20 and 70%) to species with low glass transition temperatures. The low glass transition temperature can see the ARC swell or melt compromising the fidelity of pattern reproduction. Highly cross-linked polysilanes are more immune to this phenomenon and are thought as such to even behave well as hard masks.

Antireflective coatings adjacent to the photoresist can also be implicated in LER propagation. Foot generation either in photoresist etching (trimming) or foot generation in ARC etching, if not cleared can lead to thickness variations that can represent themselves as roughness in ion driven etch processes. Furthermore, if plasma driven processes compromise the integrity of the ARC, as described by Sato et al.[31] for example by changing the film mechanical properties (T_g changes correlate with elastic properties), then this may result in sidewall contour variations.

An alternative to trim is notched gate engineering described by Foucher et al.[37] They describe taking advantage of a sidewall passivating anisotropic main etch process that is followed by an unpassivating, reasonably fast isotropic etch process that could contain, for example, SF_6. This would allow the well organized undercut to trim or reduce the gate width just above the gate dielectric by permitting the isotropic etch process in soft landing (to be discussed later) to laterally etch or undercut the passivation-free walls.

21.3.2.2 Main Gate Electrode Etch

The main gate electrode etch is intended to provide essentially vertical profiles without microtrenching. Profile variations may be induced by the etch process itself (polymerization or ion directionality) or by variations in the properties of the silicon structure with etch depth (doping concentration, local morphology). In polysilicon etch, chlorine is typically combined with additives to mitigate microtrenching and to tighten the sidewall profile. An additive or primary Si etchant, HBr promotes polymerization on polysilicon sidewalls and oxygen provides a thin oxide layer, which inhibits chlorine attack on Si sidewalls. To illustrate the nuance involved in the main etch that follows the primary patterning steps, consider the $Cl_2/HBr/O_2$ process described by Tuda et al.[38] in which a 250 nm *p*-doped polysilicon layer is etched with an oxide hard mask over thin oxide film. In their ECR system, they observed that selectivity between polysilicon and SiO_2 was greatest for the largest HBr concentrations used. However, they saw that there was a complicated shift in profile from tapered to vertical to another taper with a shift in minimum gate structure thickness (minimum CD) as the process was varied from dilute to HBr rich. Critical dimension shifts are typically associated with shadowing, i.e., closely spaced lines see less reactive or inhibiting species reaching the horizontal space in between lines because the aspect ratio between them gets larger as etching proceeds. Inversely, etch inhibitor generated on the surfaces being etched has a higher probability of being captured by the sidewalls for more closely spaced lines. In Tuda's experiments, taper in closely spaced features was found to be due to etch inhibitors generated from the surface; taper in isolated features with a large open area is the result of deposition from etch inhibitors from the plasma. Before and after-clean, views of a gate profiles etched in the HBr/O_2 plasma in Figure 21.13 show the tapered profiles that result from thick polymer deposition. The uppermost feature in Figure 21.13 has relatively vertical sidewalls but this is seen to be due to a thick polymer coating as the final structure has significantly sloped sidewalls (lower feature) when cleaned.

For sidewall profile control mediated by polymerization, CF_4 addition would at first seem to be a poor choice given the poor level of polymerization in CF_4 plasmas. Indeed, addition of CF_4 to gate etch processes is referred to as a "self-clean chemistry."[39] Polymerization is mediated by CF_4 and HBr when H (from HBr) leaches F (from CF_4) from the bulk plasma leading to HF production. More polymerizing conditions exist within trenches being etched. Xu et al.[40] describe the behavior of this chemistry in an

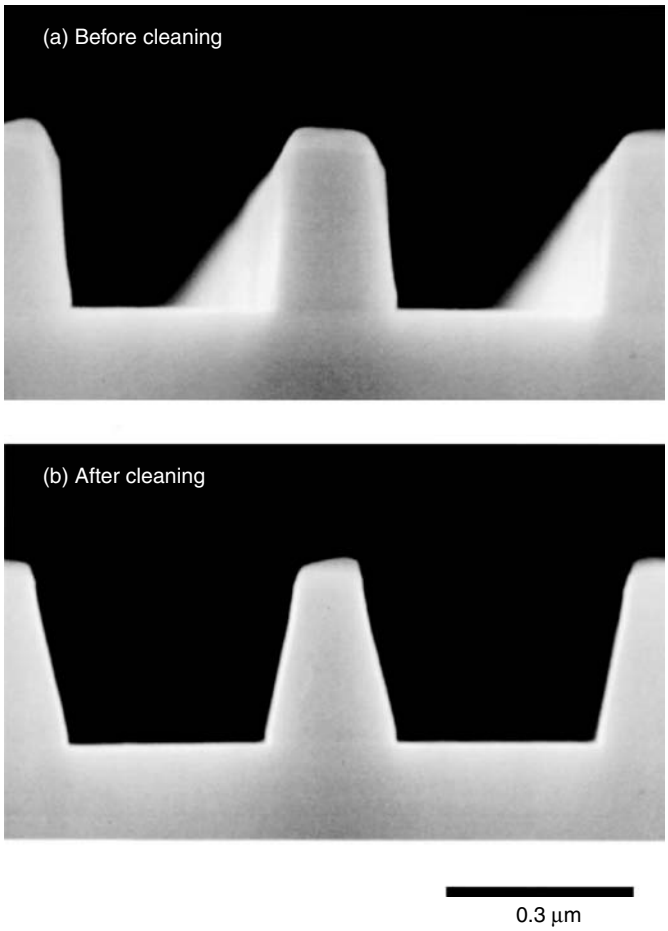


FIGURE 21.13 Impact of polymer on ultimate profile of polysilicon gate. The top structure: (a) is prior to clean. The bottom structure and (b) is post-clean hence with sidewall polymer removal. (Reprinted from Tuda, M., Shintani, K., and Ootera, H., *J. Vac. Sci. Technol. A*, 19, 711, 2001. With permission. Copyright 2001, American Institute of Physics.)

Applied Materials, Inc (AMAT) DPS source and note that O_2 addition can control the polymer buildup thickness on the sidewall. Oxygen addition, as Xu et al. observe, produces volatile COF and COF_2 products leaching CF_x from the sidewalls making them in turn less chemically etchable. A negative impact of CF_4 addition is that this sidewall polymerization can place an additional burden on the post-etch clean process as polymers coat the chamber wall as well. This makes the O_2 flow a critical parameter during the polysilicon etch process. Vallier et al.[39] show the important relationship between the chamber walls in high density plasma sources and etch processes occurring on the wafer for processes that include O_2 and CF_4. Their x-ray photoelectron spectroscopy (XPS) measurements show that the fluorine from CF_4 can generate volatile etch products from the chamber sidewall (SiOF) that may be re-cycled to the silicon sidewalls. SiOF with CF_x from the plasma are hypothesized to combine to form the sidewall passivation layer.

A complication to etching device wafers is that *n*- and *p*-doped polysilicon etch differently from undoped polysilicon. This is a concern both with respect to lateral and vertical etch rates. In general, *n*-doped polysilicon etches faster than *p*-doped polysilicon, which etches still faster than undoped

polysilicon. Addition of CF_4 is known to minimize this *n*- and *p*-type etch rate bias.[39] This is presumably because the sidewalls are effectively masked from etching radicals and, as in dielectric etching, the polymer layer controls the kinetics of the etch process. The exact mechanism is not clear. Hung et al.[41] have demonstrated process windows for matched *n*-type, *p*-type, and undoped polysilicon etch on patterned wafers without recourse to CF_4 using a Cl_2, HBr, O_2 ICP source. Specifically, they employ a LAM TCP source with 7–14 mTorr pressure, less than 120 W wafer bias power and low inductive coil powers ($<$350 W) and 10–50 sccm Cl_2 flow rates.

21.3.2.3 Gate Etch Soft-Landing and Overetch

In order to clear the polysilicon from the gate dielectric below it without detrimentally affecting the gate profile, overetch, and soft landing steps are usually required. By virtue of shadowing at the sidewall bottom, footers may evolve that cannot be removed with the main etch while retaining the oxide below the Si. Ion driven etches with appropriate selectivity are required to knock-back the footers, potentially aided by pressure ranges that promote scattering of ions to the corners. The overetch and finally soft-landing must retain the CD, original sidewall profile and not seriously interact with the silicon below the gate dielectric. Tuda et al. characterized an ECR based overetch process.[42] The process included HBr/O_2 without Cl_2. Negative CD shifts (through sidewall polymer growth) have been observed through the deposition of $SiBr_xO_y$ on the upper part of the polysilicon stump. This deposition can be seen as a function of overetch time in Figure 21.14. In their illustrative process, "footers" were essentially removed after 20 s but the gate top was thickened by 15 nm. Lateral etch mitigation in overetch processes with $HBr/O_2/Cl_2$ are thought to be due to the ability of oxygen to remove CF_x polymer from the sidewall, removing a source of spontaneous Si etching.[39] n–p Bias mismatch would presumably be a risk of this strategy.

O_2 can lead to oxidation of the underlying silicon and growth of the oxide layer in the overetch step. Knocking back footers is critical to device performance but if silicon below the dielectric layer is consumed to in fact grow the layer, when the dielectric is removed, the final Si surface may be below the gate. The mechanism of recess creation is illustrated in Figure 21.15. Vitale and Smith[43] characterized the oxidation of sub-surface silicon using a Deal–Grove model. They point out that the activation energy for oxidation is lowered to approximately 0.02 eV under the action of ion bombardment even during the soft-landing process. It is conceivable that energetic ions may penetrate 1–5 nm into the Si in effect catalyzing the oxidation process. Vitale and Smith arrive at a process that minimizes parameters that would lead to increased oxidation rates (shorter process time, lower source and bias powers, lower substrate temperatures and more dilute O_2 percentages) and successfully show reduced silicon recess of 1.8 nm as shown in Figure 21.16.

It is probable though that any process that includes oxygen will see transport through dielectric layers and significant sub-surface oxidation. The ion energy may not even need to be supplied by the wafer bias as high density plasma sources may capacitively couple enough at low source power to energize ions at the wafer. Emergent multi-frequency capacitively coupled sources may always have a significantly large plasma potential driven by the sustaining plasma electrode to catalyze oxidation by energetic ion bombardment. It is not noted in the literature but over-voltages at ignition may similarly drive ion heating leading to sub-surface oxidation.

21.3.2.4 Integration Issues in Gate Fabrication

Beyond profile control and selectivity, the combination of pattern layout, chamber condition and process sequence can render an otherwise effective etch process ineffective on parts of a layout (pattern density effects, iso-dense bias effects) or totally ineffective (etch stop conditions). The following sub-sections examine some of the more important integration issues.

21.3.2.4.1 Role of Chamber Walls

Whether from a build up of an energy dissipating layers or the coverage of chamber surfaces by reactant species, a significant change in the etch rates, even to the point of etch stop, may occur either on

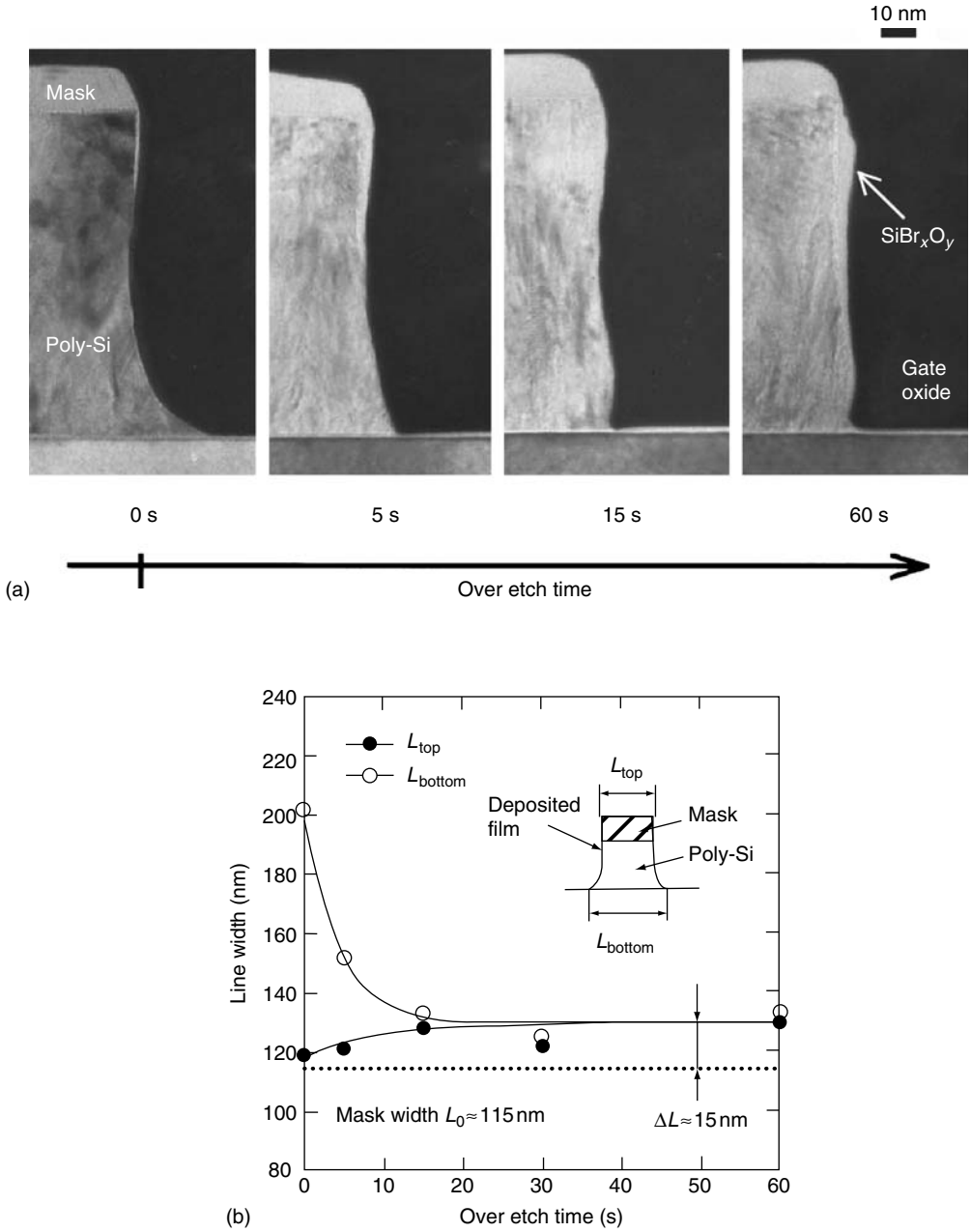


FIGURE 21.14 Evolution of polysilicon line during the overetch process: (a) feature profile and (b) line width. (Reprinted from Tuda, M., Shintani, K., and Tanimura, J., *Appl. Phys. Lett.* 79, 2535, 2001. With permission. Copyright 2001, American Institute of Physics.)

patterned surfaces or blanket surfaces. In careful XPS analysis of surfaces, Bell et al.[44] showed that silicon and oxygen from the quartz dome etched by the Cl_2 plasma resulted in the deposition of silicon oxide films on blanket surfaces eventually leading to etch stop conditions. Etching with HBr that would not attack the quartz chamber in their helicon system showed no etch stop. They noted that material from the

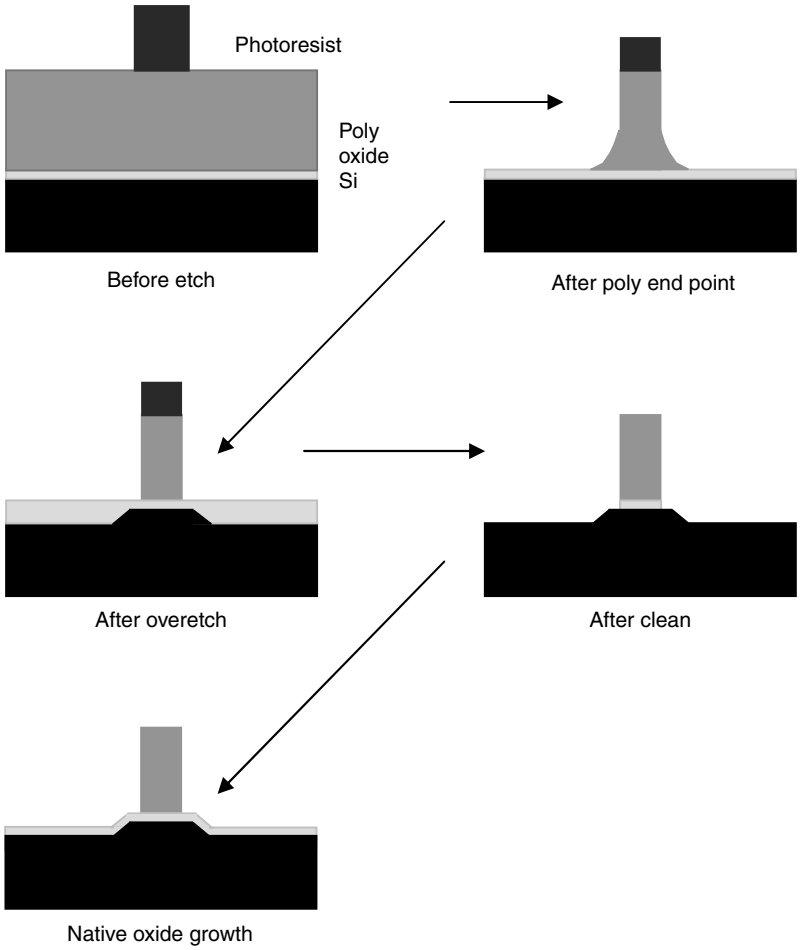


FIGURE 21.15 Recess creation through silicon oxidation and oxide removal after clean. (From Vitale, S. A. and Smith, B. A., *J. Vac. Sci. Technol. B*, 21, 2205, 2003.)

photoresist (carbon), etched material and dome material can play a significant role in gate profile anisotropy. An artifact resulting from a remote high density helicon source being employed in the study rather than an in situ source was that an oxide hard mask resulted in no appreciable oxide deposition on the sidewalls of the gates being fabricated. Why this occurs is probably related to the volatile nature of the oxide etch products being rapidly pumped away from the oxide surface. Also, the fact that the high density plasma near the quartz chamber can produce much more silicon and O when compared to the low density source near wafer in their experiment may have had some effect. If during etching, the chamber walls are brought into closer proximity to the wafer or the plasma is at higher pressure, the reflux of silicon and O_2 to the wafer governed by diffusion and flow would be more important. Bell et al.[45] observations speak to the extreme importance of chamber design on the etch unit processes and integration in general. In less aggressive etches with respect to chamber walls (e.g., HBr based etches), it is observed that oxide may be formed through the evolution of oxidized etch products that can be re-introduced to the features being evolved.[46]

Xu et al.[47] noted the impact of CF_4 and O_2 addition on sidewall profile linking the plasma source surfaces chemistry to the supply of surface species that control profile and etch rate. Specifically, they showed the influence of the gas chemistry on chamber seasoning and the seasoning impact on wafer-to-wafer effects. In

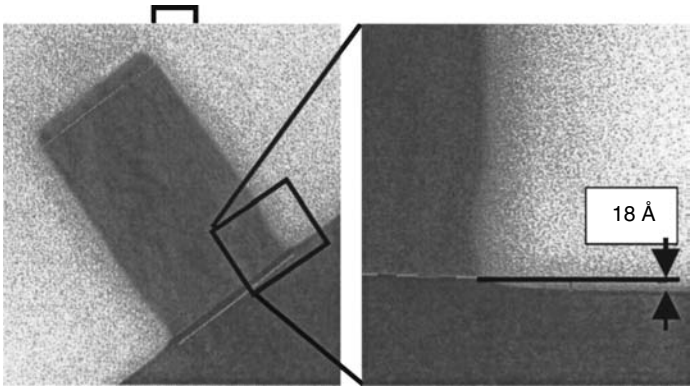


FIGURE 21.16 A minimum recess process probably limited by native oxide thickness. (Reprinted from Vitale, S. A. and Smith, B. A., *J. Vac. Sci. Technol. B*, 21, 2205, 2003. With permission. Copyright 2003, American Institute of Physics.)

an applied materials DPS chambers with anodized aluminum surfaces, the first wafer in a clean chamber had higher etch rates than subsequent wafers in both HBr and Cl_2-based plasmas with O_2. The etch rates dropped more for higher pressure, and the reduction was larger for Cl_2 plasmas. The explanation was related to the sticking rate of radicals on the aluminum surface as opposed to surfaces coated with SiO_2. When the surface was coated with SiO_2, they were more primed to absorb Cl radicals. The result is a loaded system in which the etch rate drops in proportion to the drop in radicals density due to the greater absorption rate of the large area dome. Therefore, while O_2 is important for profile control, adding more oxygen can negatively impact etch rates by promoting the coating of otherwise unabsorbing domes with absorbent material.

It is difficult to generalize the degree to which wall conditioning or cleaning is important. The aforementioned examples show that each case depends on the plasma chemistry, wall material and nature of material generated from the wafer.

21.3.2.4.2 Microtrenching Profile Generation

Microtrenching, usually related to ion scattering from the sidewall or ion attack at poorly polymerized corners, is dangerous in the context of retaining the integrity of the underlying dielectric. Bell et al.[44] showed that microtrenching is related to ion energy and could be switched off by a two step (high-low wafer bias) etch sequence. Their experiment used a remote helicon source with independent wafer biasing and Cl_2/He gas mixtures, isolating the effects of ion energy in their polymerization free process. In experiments where polymerization is more important, micro-trenching is related to the thickness of the film at the trench bottom corners with the degree of microtrenching governed by polymer precursor ion and neutral flux, ion energy and aspect ratio. There are numerous illustrative studies of microtrenching characterization and mitigation. Using feature scale modeling, Kraft et al.[48] have related ion energy conditions and reflection properties to the location of oxide punchthrough measured in experiment with modified MOS capacitors. Chemistry dependencies have been cleanly characterized by Desvoivres et al.[49] with microtrench free processes for HCl and HBr contrasting a Cl_2 plasma process with significant micro-trenching shown in Figure 21.17. In Each Case Bulk Si is Etched with a SiO_2 Hardmask Used for Patterning.

21.3.2.4.3 Uniformity

Flow and simple mass transport are known to strongly impact uniformity. As such, much effort is devoted to the design of showerheads or the design of process gas inlets and vacuum pumping. Panagopoulous[50] used as an example the design of flow in the AMAT DPSII system to describe a situation where gas is allowed to impinge the wafer at the center. Gas flow drove polymer precursors away from the center with less polymerization leading to more vertical profiles.

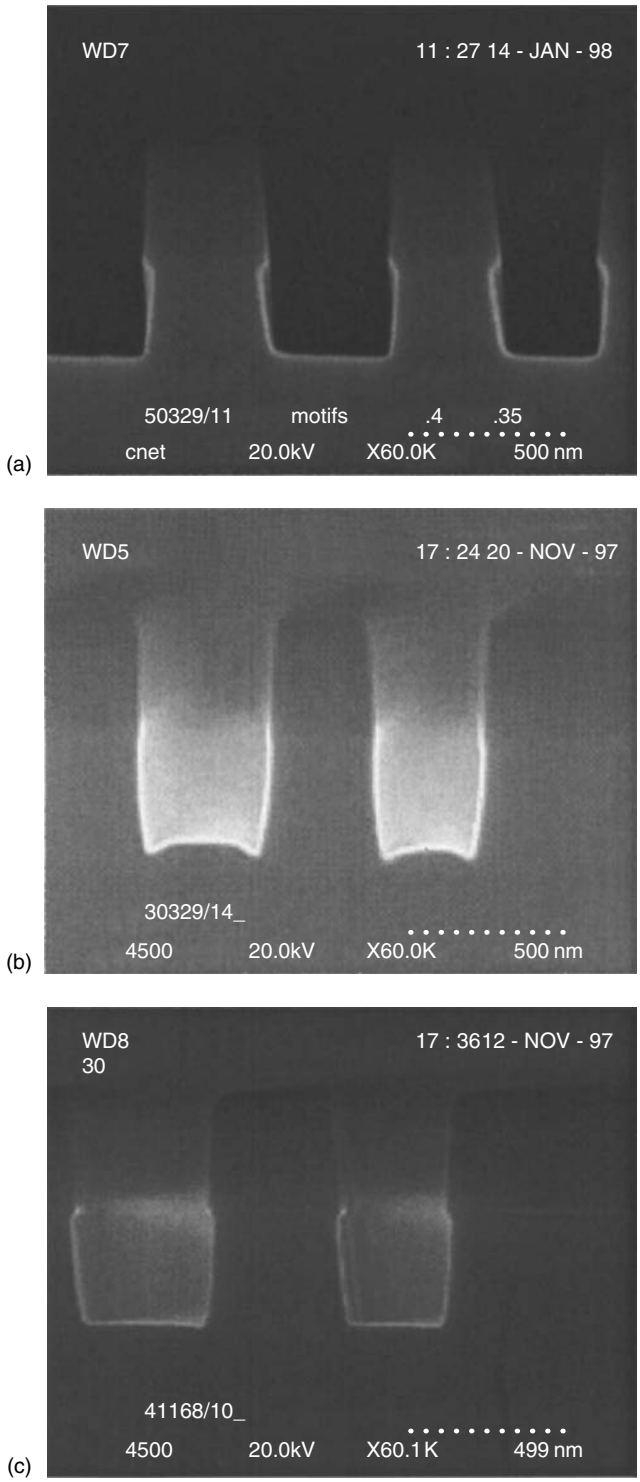


FIGURE 21.17 Effect of etching chemistry on micro-trenching in a Si etch process employing SiO_2 as a hard mask: (a) HBr, (b) Cl_2 and (c) HCl. (Reprinted from Desvoivres, L., Vallier, L. K., and Joubert, O., *J. Vac. Sci. Technol. B*, 18, 156, 2000. With permission. Copyright 2000, American Institute of Physics.)

Power coupling to the plasma and plasma interactions with materials in the chamber and wafer strongly impact etch uniformity. Managing etch uniformity is in large part a matter of managing process materials interaction at the chamber and wafer surfaces, which is often understood in terms of how the mass balance of etch precursors or inhibitors is managed in the chamber. The dielectric constant of materials surrounding the wafer impacts how power couples through them leading to variations in temperature and etch rate across the wafer, and variations in plasma uniformity.[51] Differences in plasma etch precursor reactivity between sites at the wafer edge and beyond can lead to significant differences in etch rate between the wafer edge and wafer center.

21.3.2.5 Managing Critical Dimension Biases

Critical dimension bias, defined as the variation of the minimum dimension between optical printing and the cleaned etched structure after etching, as allowed by the preceding discussion is a complicated function of the instantaneous etch chemistry, instantaneous geometry (layout), materials and other factors. Complicating matters more are that the integrated etch bias between as developed inspection (the top down imaged profile after the photoresist is developed) and as cleaned inspection (the top down profile of the polysilicon after wet clean and exposure of the underlying polysilicon) is the product of etch biases that are non-linearly related to one another and are not a linear function of feature height. On different locations of the same structure, the so-called local aspect ratio dependent effects may be different. These differences are due to each location on the structure having a different solid-angle exposure to the bulk plasma and receiving different amount of material cast at it by other features/surfaces.

The non-linear relationship between individual processes is illustrated by comparing biases one would expect for a resist trim etch step and the main etch. In the trim etch step, the lateral etch of the photoresist is primarily governed by how many oxygen radicals and ions reach the sidewall. Redeposition of material appears to play a minor role. The etch is governed by shadowing effects. If one ignores issues associated with etch resistance, mechanical buckling and LER, the lateral etch kinetics of the main etch for silicon are governed by polymeric layers of Si–O–Cl–Br–F material that build up on the sidewalls as the sidewalls are etched. As a result, complicated shapes can evolve and either negative or positive CD biases can be created. Effects such as silicon recess and punchthrough are amplified by minor tweaks in the top to bottom profile. The role of sidewall chemistry in CD budget control is presented by Desvoivres[49] for 0.1 μm patterned samples pointing out that the sidewall polymer represents a significant fraction of the CD budget.

Detter et al.[52] characterize the dependence of so-called iso-dense bias on plasma chemistry. $HBr/Cl_2/O_2$ plasmas evolve the pattern profile through the deposition of SiO_xCl_y films on sidewalls that are etch resistant. Isolated features are fully exposed to the plasma and etch resistant film builds up resulting in a tendency to taper or slope the sidewall profile. Dense features shadow the sidewall from redeposited SiO_x material and etch reaction product polymer layer. The result is a tendency to bow beneath the mask or photoresist. These tendencies are dramatically illustrated by the measured profiles in Figure 21.18. They show isolated features tapered by main etch exhibiting undercut after the overetch. In contrast, features in closer proximity are bowed at main etch and no undercut is found after overetch. The addition of CF_4 tends to thicken the sidewall layer deposit (produce taper) and reduce iso-dense bias. The combined shadowing of sidewall SiO_x of highly sticking deposition pre-cursors from the plasma and the wafer surface are responsible for top to bottom variations as the isolated features evolve in absence of other sidewall conditioning species. CF_4 chemistries by themselves are relatively non-polymerizing with the primary CF_x etch product being CF_3 which has a very low sticking coefficient. The result is that low sticking coefficient species can be transported through multiple "bounces" deep into high aspect ratio features making the polymer build-up more uniform. Variations in CD bias that would be amplified by variations in shadowing of species from the plasma through for example aspect ratio variations are minimized when low sticking coefficient species are the polymer precursors. Oxygen addition to CF_4 plasmas reintroduces chemistry and aspect ratio dependent biases by removing some of the more

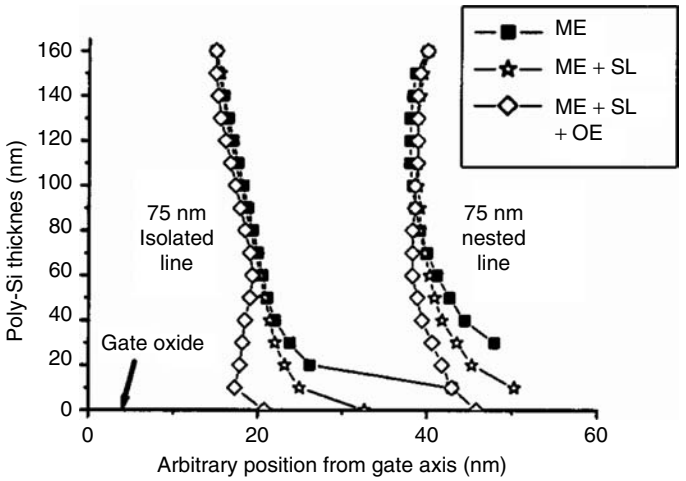


FIGURE 21.18 Gate profile after each step of the $HBr/Cl_2/O_2$ Si etch process for isolated and dense lines. (Reprinted from Detter, X., Palla, R., Tomas-Boutherin, I., Pargon, E., Cunge, G., Joubert, O., and Vallier, L., *J. Vac. Sci. Technol. B*, 21, 2174, 2003. With permission. Copyright 2003, American Institute of Physics.)

uniform sidewall polymer layer. One problem with CF_4 is that it is less compatible with oxide hard masks due to reduced selectivity. Alternative hard masks that are less sensitive to fluorine attack or more selective chemistries will therefore be needed.

21.3.3 Substrate Contact Dielectric Etch

Electrical contacts to the transistor terminals are made by etching a hole or via through the insulting dielectric that is deposited over the electrical circuit features on the substrate. This dielectric is usually chemically deposited doped silicon oxide and is usually thick to allow for planarization of the circuit features before the interconnect layers are fabricated. Because of the larger number of contacts on each IC, contact size becomes a critical factor in defining the maximum circuit density. Also, with the thick dielectric layer, contacts usually have the highest aspect ratio features that are etched during IC fabrication. Beside the difficulties associated with high aspect ratio oxide etching, additional concerns with the contact etch module include the interaction with the metallic silicide contact layer and stop layers used to fabricate more complex contact features such as SAC.

21.3.3.1 Etch of Films Deposited as Ohmic Contacts and Considerations

Silicides are deposited on gates, sources, and drains to minimize the contact resistance. The overall silicide process is reviewed by Gambino and Colgan.[53] In general, apart from polycide formation on a gate stack, the primary concern when plasma etch and polycides or salicides are involved is the process of fabricating contacts to them, i.e., etching through a material to fabricate openings that will be metallized to link to the contacts.

21.3.3.2 Contact Etch Stop Layers (SiO_2 to Si, Nitride, and Nitride to Isolation and Si)

Contact or SAC etch processes are high aspect ratio oxide etch processes that are generally selective to an etch profile guiding layer above a gate or a spacer such as Si_3N_4. The high selectivity may lead to etch stop and as such multiple step etch processes have been considered.

21.3.3.3 Contact Resistance

Conventional contact formation through an isolation layer as described by Sekine et al.[54] involves the goal of achieving high aspect ratio etch opening to active silicon, and more recently to a contact silicide or polycide, leaving conditions for good contact resistance. There are several sources of contact resistance degradation from the etch process. First, the etch process can leave contaminant. Sekine describes a CHF_3 based etch process followed by a downstream plasma clean with CF_4 and O_2. In a much earlier work, Schreyer et al.[55] describe a CHF_3/O_2 contact etch followed by a NF_3/Ar low bias power clean for use with titanium contacts. In principle, carbon based residue is removed by such a process but a second source of contamination, oxygen from the SiO_2 etch itself is enough to create a thin layer of Si–O bonds. In the case of contact to silicides and often in the case of contact to polysilicon, a plasma process by itself is not adequate to remove this contaminant as the process that will remove the layer would also supply oxygen from oxide in close proximity. Interestingly, a third source of disturbance for the silicon underlayer, amorphization by ion bombardment or radiation damage from the plasma had little impact in Sekine's results. It is often up to a dry chemical clean process of the type described to reduce contact resistance. The growth of the W plug they describe is impacted by the contaminated silicon surface but it is primarily the altering of the electronic nature of silicon surface that decreases the contact resistance.

Ikeguchi et al.[56] describe similar contact degradation issues through the formation of WO_x in the contact etch when phosphorus-doped silicon glass was etched in a RIE etcher using $Ar/O_2/C_4F_6$ accompanied by O_2 ash and $NF_3/N_2/H_2$ based polymer removal. A complex cleaning process is described in their work comprising of a sequence of ash, post-etch treatment and cleans with hydroxylamine-based solvents and sulfuric acid, critical to removing the tungsten oxide.

Consider the more complex picture described by Kim et al.[57] in which a W plug connects contacts CoSi through a via with a Ti/TiN liner. While the silicidation may or may not be straightforward, etching the contact to the liner without passing through it is not trivial with a potential result shown in Figure 21.19

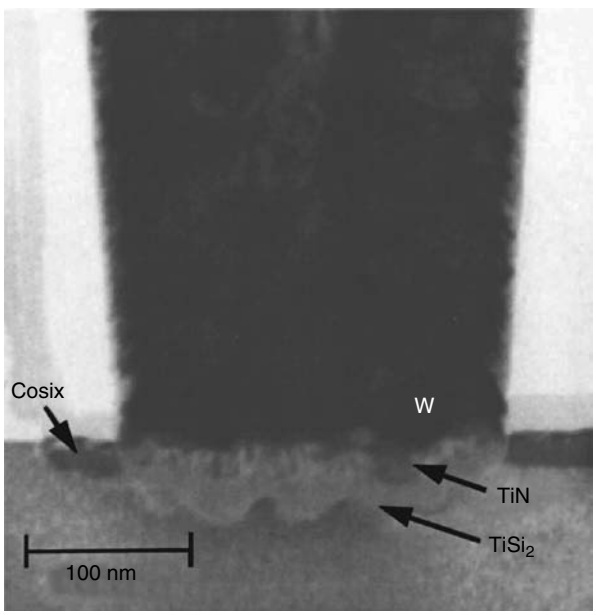


FIGURE 21.19 Polycide etch with no etch stop layer and etch through with TiSi formation. (Reprinted from Kim, J. S., Kang, W. T., Lee, W. S., Yoo, B. Y., Shin, Y. C., Kim, T. H., Lee, K. Y., Park, Y. J., and Park, J. W., *J. Vac. Sci. Technol. B*, 17, 2559, 1999. With permission. Copyright 1999, American Institute of Physics.)

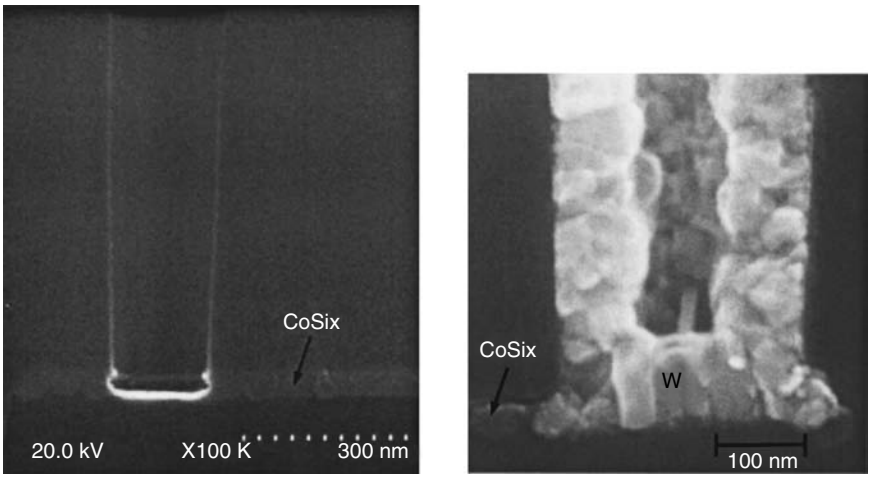


FIGURE 21.20 $CoSi_x$ etch with lateral encroachment of W. (Reprinted from Kim, J. S., Kang, W. T., Lee, W. S., Yoo, B. Y., Shin, Y. C., Kim, T. H., Lee, K. Y., Park, Y. J., and Park, J. W., *J. Vac. Sci. Technol. B*, 17, 2559, 1999. With permission. Copyright 1999, American Institute of Physics.)

in which the contact has been compromised. An etch stop layer such as SiN may be used to avoid this. Lateral erosion adjacent to the CoSi layer may result during the subsequent removal of the barrier with a wet (HF) process or the clean which can itself be a wet process. The erosion can impact coverage of the barrier and the W plug can undergo lateral encroachment illustrated in Figure 21.20. To prevent this, a "soft" very low wafer bias argon plasma clean is employed to replace the wet clean in Kim et al.'s study. This example illustrates the collision between device integration, process integration and the etch unit process via the requirements of etch stop layers and pre-metallization clean after the contact etch to bring about a successful contact module.

21.3.3.4 Aspect Ratio Etch Stop

Reactive ion etching lag, or the slowing of the etch process with increasing aspect ratio, can generally be attributed to a lowering of the ion flux reaching the bottom of the contact via due to charging effects and neutral starvation due to Knudsen transport or shadowing. The formation of fluorocarbon films at the bottom of high aspect ratio contacts is, on the other hand, the culprit for etch stop. Ohiwa et al.[58] characterized the complicated etch stop dependencies for contact etch through borophosphate silicate glass (BPSG) of varying dopant concentrations. Etch stop was only observed with high aspect ratio features and etch stop occurred at larger aspect ratios for more highly doped BPSG. In their experiments, low doping corresponded to 4.5% B and 4.2% P (by weight) and the highly doped film included 5.5% B and 6% P. The variation of etch depth with degree of doping negated the influence of current loss as a cause of etch stop. Instead they showed that etch stop is related to the transport of energetic neutral species in the feature. Fluorocarbon and inert Ar ions from the c-C_4F_8/Ar/CO_2 plasma bombard the sidewall both creating and sputtering the polymer layer. Etch products (Si_xF_y) are also deposited on the sidewall. It is probable that carbon reacts with the etch products that mostly are deposited on the sidewall with increasing aspect ratio to form SiC. SiC can itself then be redeposited at the bottom of the contact hole leading to etch stop. Higher etch rates with more lightly doped BPSG adds more oxygen to the dynamic in the via, which reduces the propensity to etch stop as the oxygen volitalizes the carbon species.

Generally true for many types of integration, the dynamics of contact etch underscore the importance of etch profile on etch process robustness. Typically more polymerizing and more selective processes will tend to sloped sidewalls. It is the interaction with sloped sidewalls both from the perspective of etched material and incoming ions and radicals that can tip the process to etch stop conditions.

21.3.3.5 Process Window

Related to etch stop is the process window for selectivity. As in the case of BEOL etch processes in which a dielectric (SiO_2 or a low-κ material) must land on an etch stop layer (conventionally SiN_x, SiC_xN_y etc.), it has been shown that selectivity of silicon to SiO_2 is dependent on the so-called polymer layer thickness. Selectivity is largely related to the polymer layer thickness for one etch process relative to the next. Put aside are polymer layer composition arguments relevant to etch stop conditions.

To obtain very high selectivity means to delicately control the thickness of polymer layers with a balance of etch and deposition between 1 and 4 nm. With highly polymerizing plasmas composed of c-C_4F_8 for example, conditions that would maintain a thin polymer layer in the etch of SiO_2 over a thick polymer layer process through silicon are limited. For a dual frequency plasma source, Matsui et al.[59] report that the thickness of the polymer layer thins for high aspect ratio contact holes (on the order of an aspect ratio of 10) and that the film themselves become carbon rich. Increasing the carbon content of films drops the etch rate, more than would be the case, through simple aspect ratio dependence and concurrently the selectivity drops. In highly dilute c-C_4F_8 plasma with 400 sccm Ar and 8 sccm O_2 at 30 mTorr, Matsui et al. showed a window of high selectivity to Si_3N_4 and silicon existed only for flow rates from 8 to 15 sccm c-C_4F_8. While this is a snapshot in parameter space for their experiment and related to the low frequency bias that they applied at the wafer (0.8 MHz), their experiments reveal the highly coupled nature of the contact etch process. Maxima in etch rate can exist over very small flow ranges and do shift with aspect ratio.

In general, high selectivity is reported for high carbon to fluorine ratio process gases: C_2F_4, C_2F_6, and C_3F_6 with H_2 rather than CHF_3 or CF_4. The rationale for the use of additives to achieve selectivity (already stated argument in gate etch) is that the gas phase becomes richer in polymer forming precursor species. All selectivity arguments retreat to the nature (thickness and chemical impedance) of the polymer layer.[60] There is a twist to the selectivity discussion when the ion trajectories are not normal to the surface being processed. The polymer layer thickness still controls the selectivity but as ion etch yields tend to be higher at off-normal angles of incidence, selectivity can be lost for highly selective etch processes as the polymer layer can be significantly eroded under those conditions.[61] This is relevant to the etching of SACs.

21.3.3.6 Self Aligned Contact and Selectivity

Self-aligned contact etch or borderless etch processes allow a contact etch to land on both the source/drain region and gate region simultaneously or source/drain and isolation region simultaneously (silicide and oxide). These processes can significantly increase circuit layout density and relax the lithography requirements but they impose tough selectivity requirements on the etch process as multiple materials are landed upon in one process. As described by Givens et al.[62] the process window of the SAC process is many dimensional with each factor highly coupled to the other. In the conventional process, the contacts are made through doped SiO_2 and exposed to a conformal silicon nitride (or similar insulating barrier) film. This is followed by a nitride etch step that should not etch deep into silicon or oxide isolation and silicide regions. In addition to typical process parameters, Givens and co-workers showed significant equipment dependence of the etch process. Moving from a capacitively coupled system to a high density plasma source with C_2F_6 resulted in significant increase in selectivity; and for their structures, significant reduction of the aspect ratio dependent etch effects or etch stop. Broadly, one can now interpret their results as allowing for much higher fluxes of radicals moving the etch process to the ion starved regime where the directional ion flux sets the etch rate. All reported aspect ratios were well below that discussed by Ohiwa et al.[58] and Matsui et al.[59] so that ARDE would not be a factor in this regime.

Givens' early conclusions regarding etch rates and profiles include trends that are true even today for highly polymerizing, high density plasma etch of oxide and nitride materials. Increasing flow rates and pressure (all other parameters constant) tend to produce higher etch rates. This is essentially due to the delivery of more etch precursor ions or radicals. Increased pressure also leads to the delivery of more neutral species that can polymerize on surfaces thereby leading to profile taper. Increasing the source power decreases the etch rates due to a decrease in sheath potential at the wafer for higher plasma densities and plasma currents. Conversely, increasing the bias power at the wafer increases the etch rates. Increasing both powers makes sidewalls more vertical as they are cleaned of protective polymer layer that would lead to taper.

The interplay between plasma chemistry and aspect ratio can have a serious impact on topography. Shon et al.[63] and Kim et al.[64] show that polymerization can lead to the formation of steps on nitride in SAC etch and in essence modulate selectivity. Shon describes of ion precursors that can lead to polymerization. Kim et al. show how the ratio of the exposed oxide area to the exposed nitride area can impact the Si_3N_4 etch rate and ladder formation. An example of ladder formation is shown in Figure 21.21. In addition, as in gate etch, SAC etch is impacted by spacing. Larger contacts (contact area relative to spacer nitride cap) may result in oxygen (from SiO_2) eroding the plasma produced polymer providing the selectivity to silicon nitride. For the same process, larger contacts may therefore etch less selectively than smaller ones.

Spanning the range of low to high selectivity with a Freon based SAC process, Quiao et al.[65] show that poorly selective processes can lead to punchthrough of the nitride layer or, in a slightly more selective scenario, "shearing" of the nitride overlayer. With overly selective processes approaching 40:1 oxide to nitride selectivity, the sidewall can become tapered leading to poor contact or no contact due to "grass" formation. A span of SAC etch process consequences from completely unselective to highly selective are illustrated using Quiao's scanning electron micrographs (SEM) in Figure 21.22. Figure 21.22a shows that

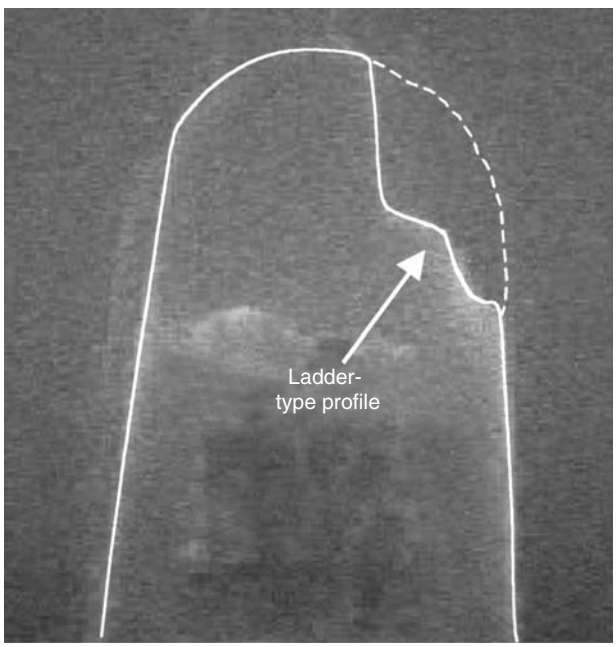


FIGURE 21.21 Ladder formation in self-aligned contact etch. (Reprinted from Kim, J., Chu, C. W., Kang, C. J., Han, W. S., and Moon, J. T., *J. Vac. Sci. Technol. B*, 20, 2065, 2002. With permission. Copyright 2002, American Institute of Physics.)

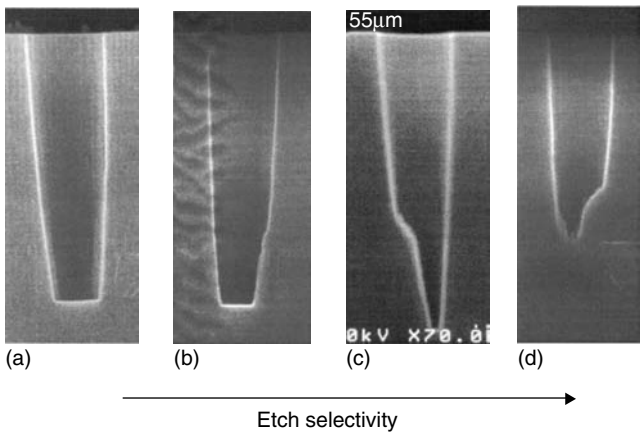


FIGURE 21.22 Variations in etch selectivity illustrating compromise of the spacer to etch stop: (a) represents a non-selective process in which the spacer is compromised, (b) and (c) represent more selective processes whereas (d) illustrates etch stop. (Reprinted from Qiao, J., Jin, B., Phatak, P., Yu, J., and Geha, S., *J. Vac. Sci. Technol. B*, 17, 2373, 1999. With permission. Copyright 1999, American Institute of Physics.)

the spacer has been completely compromised; Figure 21.22b is less compromised with Figure 21.22c seeing the contact opened. Figure 21.22d is a process that is selective to the nitride but as the aspect ratio of the feature to be etched becomes large enough, the process slows to etch stop. Behaviors ranging from lateral punchthrough of the nitride to etch stop are shown to result for three sccm variation of c-C_4F_8 flow by Kim et al.[66] and is shown in Figure 21.23. In Figure 21.23a, the silicon nitride cap is etched through (punchthrough) due to lack of a protective plasma chemistry. Increasing amounts of c-C_4F_8 result in better selectivity by small flow changes (only 3 sccm) make the difference between a highly selective process and etch stop.

Omnipresent in issues related to selectivity are wafer temperature effects. Quiao et al.[65] showed in their study that the uniformity of He backside cooling correlated strongly with selectivity. Since less effective cooling at the wafer periphery leads to greater degrees of polymer formation, minimizing the across wafer temperature differential was very important from the perspective of maximizing yield.

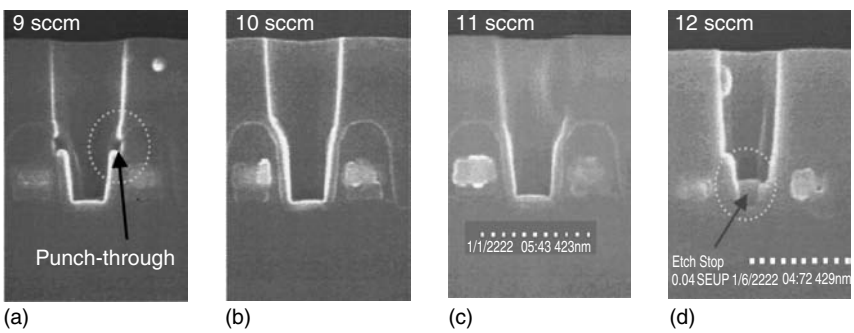


FIGURE 21.23 Impact of c-C_4F_8 flow on the transition between punchthrough and etch stop. (a) through (d) represent 3 sccm increments in the c-C_4F_8 flow with (a) being leanest in c-C_4F_8. (Reprinted from Kim, J.-H., Yu, J.-S., Ryu, C.-K., Oh, S.-J., Kim, S.-B., Kim, J.-W., Hwang, J.-M., Lee, S.-Y., and Kouichiro, I., *J. Vac. Sci. Technol. A*, 18, 1401, 2000. With permission. Copyright 2000, American Institute of Physics.)

It is reasonably well accepted that the polymer thickness controls the oxide to nitride etch selectivity. Matsui et al.[59] have correlated the dependencies of polymer thickness with O_2 content in the feedstock gas of the plasma quite thoroughly. Selectivity windows therefore can be modulated by O_2 addition. One potential issue is that, with high aspect ratio features, ARDE effects may limit the ability of O radicals to control selectivity in deep features. Alternatively, Sun and Huang[67] have proposed that the addition of hydrogen containing precursors such as CHF_3 would create weaker polymers that, while being protective, would be less susceptible to etch stop with ultra-selective C_5F_8 processes due to weaker C–H bonds.

Selectivity and its relationship to the ratio of thickness of dielectric and etch stop layers can have other implications. It is more and more being appreciated that the stress in films like etch stop layers can propagate to the gate and impact device performance. This stress is directly related to the etch stop thickness and other geometric parameters related to selectivity requirements.[68] Kim et al.[66] summarize this issue and others noting that:

- Low-pressure chemical vapor deposition (LPCVD) nitride films that have been conventional for SAC are highly tensile,
- Conventional nitride barriers have relatively high dielectric constants, and
- LPCVD nitride films have a very high thermal budget.

Replacement of the classic nitride etch stop layer is well underway with one example being SiO_xN_y. These films are a candidate because of their superior ARC characteristics, stress, and etch characteristics. Kim et al.[66] describe a process where $CF_4/Ar/O_2$ is used to pattern the hardmask and ARC in Applied Materials MxP+ while oxide is selective etched to SiO_xN_y using c-$C_4F_8/CH_2F_2/Ar$ in a Tokyo Electron Ltd. (TEL) design-rule manual (DRM) chamber. Here again, the CH_2F_2 is included in the chemistry to prevent etch stop.

In advanced applications contact will be made to a silicide deposited on Si. The etch process in this case should be selective to the silicide. When the etch is complete, a reside will be present on the silicide after etch and strip of the photoresist if it is used. This must be cleaned by a process that will not cause loss of the silicide or increase contact resistance. As was described in the earlier portion of this chapter, etch processes can alter the stoichiometry of substrates. This can happen when a plasma contacts the silicide as there may be a tendency to leach silicon from the material leaving metal rich surfaces behind. The metal rich surface can be more easily etched and potentially removed in a clean.[53]

21.3.4 Damage Considerations

It does not require a stretch of the imagination to believe that plasma processes can impart damage to electrical devices being processed. In general, the damage is related to a degradation of the electrical performance (usually a transistor) through the introduction of defects that compromise the electrical strength of the oxide or through the introduction of defects at surfaces or in the bulk that introduce charge or enhance leakage. The source of the defects can be ion bombardment, current, or photon bombardment. Damage related process anomalies (e.g., during overetch) are primarily caused by so-called electron shading, physical damage and plasma non-uniformity.

Damage is detected by a variety of electrical tests. The processes related to gate etch that are indicted in cases of damage are STI, contact etch or PECVD deposition. We will limit our discussion to etch specific damage issues.

There are generalizations regarding damage that are relevant to the devices being processed. For example, a rule of thumb is that P-channel metal-oxide semiconductor (transistor) (PMOS) devices are more prone to damage than N-channel metal-oxide semiconductor (transistor) (NMOS) devices. In the case of antenna structures, damage is correlated with larger antenna ratios (more current collection area). Furthermore, the nature of the damage is typically related to the oxide thickness. Thicker oxides (>6 nm) damage is related to defect density and detectable using threshold voltage shift or hot carrier injection measurements while thinner oxides (<4 nm) fail due to integrity loss and are less likely to fail under conditions meant to observe V_t shift and leakage.[69]

Brozek et al.[70] looked at failure through intermediate oxide thicknesses (4–9 nm) with Fowler–Nordheim electron injection and found that damage related to oxide etch was greater for NMOS structures and independent of thickness. Damage related to PECVD was greater for PMOS. Using yield as a measure, the peak damage occurred for devices with 5–7 nm oxide thicknesses.

The primary damage induction mechanisms in etch are related to plasma non-uniformity, gross process damage, and electron shadowing. When the plasma is non-uniform, the plasma potential varies across the wafer resulting in lateral current flow in the devices. Gross damage refers to the compromise of the gate oxide and recess into silicon leading to loss of device integrity. Finally, electron shading is related to the fact that electrons have a broad angular distribution while ions enter a feature anisotropically. The electrons tend to be captured by the upper sidewalls of features and the ions by the base of features. The result is that potentials evolve in features that tend to deflect ions to sidewalls. These large potentials can themselves compromise the gate oxide.

Relating damage to process conditions has historically been a post-mortem exercise. That is to say damage is induced in special test structures and the damage measured and process optimized. Recommendations related to processes (and equipment) tend to be less generally applied and more local to the exact experiment.

Chang et al.[71] for example, looked at process plasma chemistry (SF_6, HBr, Cl_2) dependencies in an ECR system. Evaluating processes using flatband voltage shift, interface state density changes and device yield, damage was associated with long overetch times and large antenna ratios. SF_6 etching conditions resulted in both plasma non-uniformity and compromise of the oxide leading to leakage. The ECR system that was used could also have non-uniformity dialed in through the magnetic coil settings to introduce damage.

Most of the time, damage is related to design parameters. Tsui et al.[72] measured low field gate current (I_g) after contact etch as a function of number and contact hole size and spacing. Damage correlated with increasing contact number and decreasing contact hole dimension. Observing essentially no NMOS variation, it was noted that usually BEOL processes annealed out damage, rendering it "latent" showing up in PMOS devices only with Fowler–Nordheim stressing.

There exists the general thought that RIE sources create more "damage" and high density plasma sources create less damage. Hashimoto et al.[73] have explored damage creation by four different high density plasma sources using special electron shaded antennas. Damage was observed in all their experiments. The damage decreased with decreasing source power. They uncovered a counter-intuitive dependence on wafer bias power in that the damage decreased with increasing wafer bias power. While they observed that more work was needed to understand the effect, it was thought that more directed electron current was drawn to the feature. That is, the bias power heats the electrons rendering the shading by the antenna structure less effective. Physical damage by etch to underlying silicon is potentially serious when there are processes without an etch stop later as in the case of STI.[74] In both Cl_2/N_2 and Cl_2/HBr plasmas used to etch silicon, defects were left in the silicon several nanometer below the surface. In principle, these defects should be removed by thermal oxidation of the surface in which 10–20 nm of silicon could be consumed. However, it has been observed that damage remains after oxidation and significant high temperature anneal is required to ultimately eliminate the defects.

21.3.5 Back End of Line Etch Processes and Dual Inlaid Processing

Back end of line etch has come to be synonymous with single inlaid, dual inlaid or damascene processing of the trench via interconnect into which (typically) copper wires are fabricated. Inlaid etches are positive processes as the photoresist is opened to where the metal is to be deposited; Al and Cu–Al alloy interconnect are negative processes in which the metal is patterned and the open areas of photoresist are etched away. Fundamentally, positive processes were needed for Cu processing as Cu readily forms precipitates with typical metal etch gases (e.g., Cl_2). While it is possible to fabricate vias and trenches at each level with an individual sequence of photoresist development and etching, cost imperatives have resulted in the development of process integrations in which multiple etch steps can be executed in sequence to fabricate both lines and via interconnections in a single module. The two most common

configurations are "via first-trench last" (VFTL) and "trench first-via last", with various permutations of these combinations in use. An excellent overview of the details of dual inlaid BEOL processing is the review by Kiel et al.[75]

The more common dual inlaid process, VFTL, is now described and a typical sequence of steps involved in VFTL are illustrated in Figure 21.24. In one sequence, with some imaging configuration, a via is patterned for subsequent etch. The details of the imaging configuration are related to the ability of one, to image the via with appropriate resolution. A single layer, bi-layer or tri-layer resist configuration can be used. Single layer resists are patterned through illumination and dissolution. Bi-layer resists are patterned through the use of an image layer and an "under-layer" which is etched using the image layer as a template and behaves as a hard mask. A tri-layer resist could be a photoresist layer beneath which is an ARC and a hardmask layer. The etch of a via through tetraethylorthosilicate (TEOS), fluorine-doped silicate glass or SiCOH is carried out with typical dielectric etch plasma chemistries in inductive or capacitively coupled plasma sources, although capacitive systems are becoming more popular for this application. An etch stop layer is usually between the via and metal levels so that the via etch can be stopped appropriately. A trench etch follows which raises the issue of stopping without eroding the already formed via. This is typically achieved by filling the via with a "slug" which can be an ARC material or photoresist itself etched so that the slug fills the via to a desired height. What the desired height is a complex function of the etch rates of the dielectric, the slug material and photoresist after patterning the trench. A trench etch erodes the remaining dielectric and the slug protects the via. The photoresist and ARC are stripped, the etch stop layer opened in another etch step and contaminant removed in a post-etch treatment.

Profile issues in via etch include the generation of striations and taper to bowed profiles. Striations are the consequence of coherent azimuthal scattering in vias as the etch front progresses, roughness templating or a combination of both. In trench etch, the generation of microtrenches related to ion scattering off of the sidewall or non uniform polymer deposition on the base of the trench can be an important issue. Most important is the integrity of the via-trench interface. If the slug material is recessed too low in the via and the via slug etch rates progresses such that the trench base never meets the level of the slug, the trench via interface erodes resulting in formation of shelves or facets. Specific plasma chemistries may contribute to this as well through deficient polymerization. An example of compromise of the trench via interface through faceting is shown in Figure 21.25a. On the other had, if the slug material remains too high in the via and the via slug etch rates progresses such that the trench base is lower than the level of the slug, the trench via interface is shadowed such that the trench via interface is not etched. The result is the creation of fence like structures around via periphery. Fences are also caused by enhanced polymerization, an example of which is shown in Figure 21.25b. Fences and facets can also be formed simultaneously, as shown in Figure 21.26.[76] The formation of both shelves and fences have implications for subsequent metallization processes, reliability and yield.

Materials interactions in etch have impact on both profile control and on materials property integrity. Slug and etch product interactions can change the properties of the material being etched altering the material selectivity. Energetic etch products and species from the plasma can penetrate the sidewall of

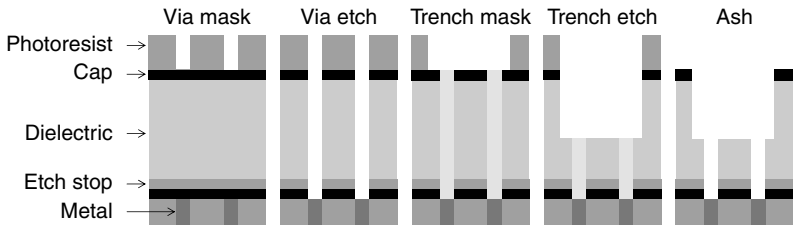


FIGURE 21.24 Steps in a typical via first trench last dual damascene process.

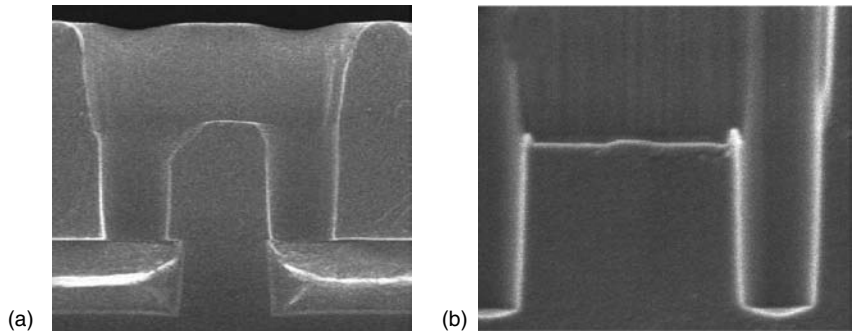


FIGURE 21.25 Dual-inlaid etch process with (a) less polymerizing and (b) more polymerizing etch plasma chemistry used in the trench etch step. The less polymerizing process leads to faceting and compromise of the trench via interface whereas the more polymerizing process leads to fence formation.

porous low-κ materials changing the dielectric constant. Surprisingly, it appears that the penetration of etch precursors into low-κ dielectrics is not a serious issue as sidewall polymer acts as a diffusion barrier. When the sidewall polymer is cleaned in a post-etch treatment or the photoresist is ashed away, material can penetrate into the now open dielectric. Currently, it is the maintenance of the materials integrity in the ash, post-etch treatments that are the major issue in low-κ dual inlaid etch module development.

While not strictly in the realm of etch, the interface of etch with subsequent metallization establishes the metrics for the dual inlaid etch process. Metallization processes that are typically PVD based tend to favor via and trench profiles that are tapered. Bowed and purely vertical profiles are very difficult to coat with a diffusion barrier. Tapered profiles tend to be amenable to being able to deposit diffusion barriers of critical thickness. In the same sense fenced profiles tend to be harder to metallize and faceted profiles are easier to metallize. Breaking through an etch stop layer can cause metal from the line below to be resputtered to the dielectric sidewall also compromising the dielectric materials integrity. Management of breakthrough and contact with underlying metal layers requires complex etch-metallization integrations.

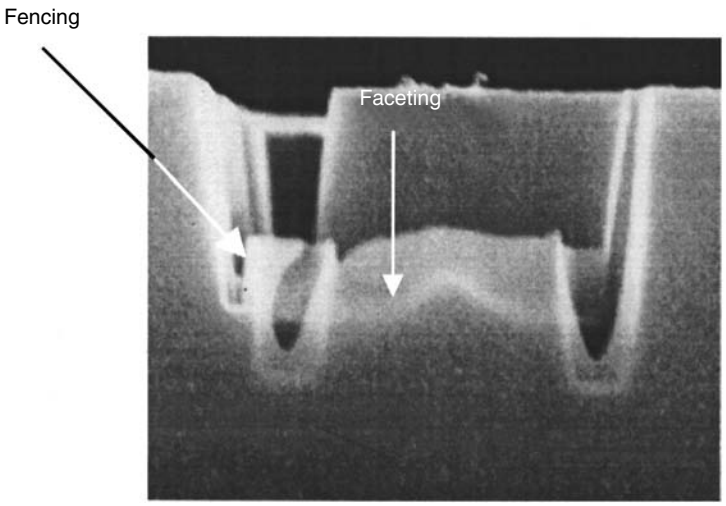


FIGURE 21.26 Scanning electron micrographs (SEM) image of via first dual damascene trench etch illustrating fencing and faceting around the via hole. (Reprinted from Kropewnicki, T., Doan, K., Tang, B., and Björkman, C., *J. Vac. Sci. Technol. A*, 19, 1384, 2001. With permission. Copyright 2001, American Institute of Physics.)

In the following sub-sections, given the complexity of etch gas and material combinations with the advent of porous low-κ materials, we start with a detailed survey of dielectric etch mechanisms and then discuss specific dielectric etch process management issues.

21.3.5.1 Oxide Etch Fundamentals and Origins of Selectivity

The phenomenology of CF_4/O_2 etching of Si/SiO_2 was initially described by Mogab, Adams, and Flamm.[77] This work showed that the etch rate of blanket SiO_2 was proportional to the fluorine concentration with oxygen playing the role of diluent. Oxide etching is today typically associated with much more complex plasmas than CF_4/O_2 including highly selective highly polymerizing feed gases (e.g., CHF_3, *c*-C_4F_8, C_4F_6, C_5F_8). Tatsumi et al.[78] Matsui et al.[79], and Kurihara et al.[80] showed a well behaved relationship between incident F radical flux, ion flux, fluorocarbon layer thickness, reactive later thickness with etch rate and selectivity. They showed that etch rate correlates with incident F flux (either as F radicals, ions, or CF_x species) through a "sticking coefficient" that depends on ion energy and fluorocarbon layer thickness. The fluorocarbon layer thickness essentially attenuates the energy of incident ions. Etch rates, known in general to correlate with ion energy, rely on energy making it through the fluorocarbon layer to drive reactive ion etch processes. High ion energy results in both more energy making it through the layer (more etching) and thinner layers (facilitating less attenuation). Low ion energies result in less energy transfer through the polymer layers and thicker films, themselves leading to more energy attenuation. It is the confounding affect of ion energy and film thickness that leads to etch stop. Overall, in the presence of a layer containing carbon, it is the net delivery of fluorine to the subsurface that determines the etch rate.

A few specific observations from the work of Tatsumi, Mastui, Kurihara, and others provide useful guidance regarding practical etch engineering. Polymer layer (C_xF_y) thicknesses for C_xF_y plasma chemistries in which etching occurs are 5–6 nm on Si and Si_3N_4 whereas it is less than 1 nm on SiO_2. The reaction layer (SiF_xO_y) is 1–5 nm on Si and Si_3N_4 but less than 1 nm on SiO_2. The polymer layer is important as it has been shown that it can attenuate the energy of incident ions by 100s of eV. The existence and thickness of the polymer or "inhibitor" layer is required to produce selectivity to underlying materials with C_xF_y based plasma processes. With polymer forming plasma chemistries, therefore, very high ion energies are required to keep the polymer layer at a critically low thickness whilst allowing the polymer layer thicknesses to be thick enough in the underlying material on which the process needs to stop. These requirements pose an inherent problem for high density plasma sources in which non-sidewall polymerization rates are low due to the high degree of dissociation of the feed gas in these plasmas and self-bias voltages are typically low compared to lower plasma density capacitively coupled sources. Polymer precursors need not only come from the gas phase[81] but may arise from any surface in the chamber. In some systems it may be that the walls are the primary source of materials such as CF_2.

Returning to the conclusions of Tatsumi et al.[78] the key to etch rates on SiO_2 is fluorine delivery, be it by direct fluorine delivery or defluorination of the fluorocarbon film. What fluorine reaches the surface, at steady state must leave the surface.[82] The thickness of the fluorocarbon and reactive layers are such that this state is achieved. Measurements on (CHF_3, CF_4, *c*-C_4F_8) high density plasmas etching SiO_2 and other substrates show that the polymer layer thickness does indeed become smaller with ion energy. However, at very high ion energies, etch rates increase while the polymer layer remains the same thickness. For CF_4 plasmas, the C_xF_y film is so thin that the fluorine is directly delivered to the substrate through the film. In most other plasmas, e.g., CHF_3 and *c*-C_4F_8, the fluorine is delivered to the film by the etch precursors and ions incident on the film, liberating free fluorine that impinges the substrate producing etching.

Additive gases can change the polymerization dynamics.[83,84] It is generally accepted that CF_2 ions result in the highest etch yield for SiO_2 and CF_2 radicals are the most depositing of CF, CF_2 and CF_3. Teii et al.[85] note that addition of H to high C/F feedstock gas plasmas terminates dangling bonds impeding or ultimately limiting polymerization in the gas phase or at the surface. Conversely, in low C/F feedstock plasmas such as CF_4 that are composed of more fluorine rich species in the gas phase and at the surface, H can abstract fluorine to produce HF leading to a more polymerizing gas phase and more polymerizing

surfaces. H_2 addition in short residence time plasmas, typical of most processing conditions, will tend to see dangling bond termination with hydrogen be dominant for all feedstock gas conditions.

One should note that Teii et al.[85] and Stoffels et al.[81] in their work use gas mixtures that tend to be more rich in etch precursor gases than is typical for etch processes. For example, Teii et al.'s Ar dilution experiments include up to 60% Ar. Stoffels' work involves pure mixtures. Teii does point out that energy transfer from excited state Ar to C_xF_y molecules and radicals can induce electron attachment and by inference other plasma processes. Given that plasma characteristics are sensitively dependent on composition, it may be dangerous to extrapolate literature conclusions about their behavior at lower degrees of dilution to higher ones. Hori and Goto[86] point out that under high density plasma conditions (and ECR plasma) it is possible to fully account for polymer deposition without recourse to gas phase polymerization mechanisms but also caution that the conclusion can be chamber and process specific.

While a general consensus exists that CF_2 radicals play a key role in etching, the exact mechanism for its role in selectivity enhancement is a matter for debate with either polymerization in the gas phase or surfaces playing a role. Use of high density plasma sources operated at low pressure results in a high degree of dissociation and a plasma made up of poorly adhering CF, CF_2, and CF_3 radicals, with a substantial quantity of fluorine, conditions that are not selective.[79,87] On the other hand, lower density plasmas operated at higher pressures with less efficient electron heating (conventionally capacitive or MERIE) will have more massive polymeric and higher carbon fraction species leading to more selective processes. A downside is that a more complex array of larger polymeric precursor can lead to polymer film non-uniformity. Lower plasma densities would generally mean that higher self-bias voltages are required to etch what can be thicker films on horizontal surfaces. All in all, etch rates and selectivity generally may be thought to be improved in capacitive systems but process windows may be more restrictive. Recent innovations with MERIE tools and dual frequency plasma sources have in part come about to address these issues. Two trends that have not quite made it to the mainstream to address these general shortcomings are ultra high frequency ICP[88] sources and neutral beam plasmas for etching. In the former, the idea is to narrow the electron energy distribution function such that one produces mostly CF_2 in a c-C_4F_8 or other C_xF_y etch precursor plasma. In the latter, the goal is to provide a source of energetic narrowly focused neutrals to effect etching without ions playing a key role.

21.3.5.2 Implications of Etching on Porous Dielectrics—Blanket Materials Conclusions

SiO_2 and fluorinated SiO_2 have given way to hydrogen and carbon containing dielectrics, either spun or deposited by PECVD with or without porosity, to achieve lower interconnect dielectric constant requirements. Pores are typically on the nm scale with porosities ranging up to 40%–50%. It is tempting to assume that the etch mechanisms for SiCOH, porous SiCOH and SiO_2 would be analogous. However, mass corrected etch rates (CER) do not simply scale with porosity; non-porous SiCOH can have very different etch rates from SiO_2.

Sputter rates of SiCOH in pure Ar plasmas have been reported to be 10 times those of SiO_2. Adding pores to the picture significantly complicates a description of the etch process. Regimes of distinct behavior include those delineated by process condition and degree of porosity. High ($\sim$50%) and low (up to 30%) porosity materials have different behavior with pore connectivity perhaps determining the behavior. Neutral starved, ion starved and highly polymerizing plasma conditions also have distinct behavioral differences.

Porosity can impact the etch process by increasing the surface area and decreasing the energy per unit area or fluence, which reactive ion etch processes tend to scale with. On the other hand, the topography may be such that certain porosities have angles relative to incident ions that can enhance the local etch yield. More importantly, porosity can facilitate mass transport through the pores, especially for highly porous materials where material can be transported deep into the substrate. These phenomena are described in recent work by Hua et al.,[89] Posseme et al.[90] and computational studies by Sankaran and Kushner.[91]

Hua et al.[89] reported porous low-κ material etch rates for Ar, c-C_4F_8 (ion starved) and Ar/c-C_4F_8 (neutral starved) plasmas in a high density plasma source. Consistent with the enhancement mechanisms just discussed, the Ar sputter rates increase with porosity and mass CER show slight enhancement (20–40 nm/min). Highly-polymerizing ion-starved plasmas such as pure c-C_4F_8 showed decreasing etch rates and CERs as a function of porosity but are approximately 10 times higher than the Ar sputtered rates ($\sim$100–200 nm/min). Neutral starved etch rates monotonically increase with porosity from 250 to $\sim$900 nm/min with CERs showing an initial enhancement with porosity and then a decrease with porosity. All of these results were at a fixed self-bias voltage of -125 V.

From earlier sections it was understood that SiO_2 etching in fluorocarbon gases can be described by looking at the net delivery of F to the wafer by all species. The idea is that ions crack the polymer, liberating F that passes through the polymer layer driving the etch process. Penetration of a fluorocarbon layer deep into the substrate inhibits the ability of the ions to crack the polymer adjacent to the dielectric. Porosity effectively thickens the fluorocarbon film layer in the manner thought of in conventional dielectric etch reducing the etch rate with increasing porosity. One significant difference is that penetration of the fluorocarbon layer into the bulk results in a homogenous Si–O–C–F layer beneath a surface fluorocarbon layer. In contrast, a fluorocarbon layer resides above a reactive Si–O–F layer in SiO_2 etching.[90]

In a neutral starved regime, Hua et al. describe a picture whereby a plentiful ion flux is dominating a lower polymerizing flux, which can be efficiently cracked at the surface with enhanced etching. Large degrees of porosity and porosity interconnection can lead to a decrease in CER through various mechanisms such as decreased local surface area as larger pores emerge and shadowing of the ions by submerged pores.[89]

Posseme presents the slightly more complex but not atypical plasma chemistry in which a non-polymerizing CF_4/Ar plasma is coupled with a polymerizing CH_2F_2 additive.[90] For their capacitively coupled plasma source studies they observe that for materials with reasonably similar stoichiometry, the CER drops with increasing porosity and drop to the point of potential etch stop with the addition of highly polymerizing CH_2F_2. They showed that etch rates decrease with decreasing porosity (50%–30%) for Ar/CF_4 discharges from nearly 1000 to 500 nm/min. Above 80% Ar dilution, the etch rates drop severely to less than 100 nm/min; above 5% CH_2F_2 dilution, etch stop was seen to occur. With CH_2F_2 dilution the etch rate of SiO_2 can exceed up to 40% porous SiCOH; in the case of no CH_2F_2 addition, SiO_2 etch rates are typically less than any SiCOH etch rate. Posseme notes that stoichiometry is of importance when comparing processes and that not all SiCOHs or porous dielectric have the same composition. Etch topography is generally observed to be rough in the neutral starved regime and smooth in the ion starved regime especially with etch stop.

Etching for very porous films likely does not come to steady state during an industrial process time with steady state being reached for SiO_2 etching on the order of 5 s, porous low-κ dielectric at 15% porosity reaches steady state in approximately 15 s. Fifty percentage porosity films go from an initially rapid etch rate to a much slower etch rate in greater than 1 min for the conditions Hua reported.

General conclusions about blanket porous dielectric (specifically SiCOH) etch include:

- Etching in neutral starved conditions (or higher density sources) is generally higher with increasing porosity,
- Etching in ion starved conditions (or lower density sources) are generally slower with increasing porosity,
- Inclusion of polymerizing additives or more polymerizing conditions can promote etch stop or remove porosity dependence, and
- Very high porosity and pore connectivity can contribute to a lowering of etch rate.

Given that material stoichiometry and recycle of etched material from the wafer through the plasma back to the wafer surface as polymerization or etch precursors can determine an actual process behavior, the generalizations here would depend highly on etch chamber configuration properties (pressure, gas

residence time, temperatures) and specific operational behavior (plasma potential, electron temperature, electrical waveforms).

21.3.5.3 Profile Control

While blanket etch rate phenomenology and etch selectivity is a starter for process understanding, selectivity and etch rates ultimately are related to feature profile evolution. What reaches feature sidewalls in etching is very dependent on the plasma chemistry. Reactive neutrals from the plasma tend to adhere to sidewalls as they are less likely to be sputtered by glancing angle ions or they may even be shadowed by the evolving feature topography. While all C_xF_y plasmas deposit material on sidewalls, it turns out that CF_4 plasmas result in a significant redeposition of material from the feature bottom to the sidewall, comparable to what reaches the sidewalls from the bulk plasma. To a lesser degree this happens for CHF_3 as well, and in the specific instance of Min's[92] source, nearly not at all for *c*-C_4F_8. More important than resputtered polymer ($C_xF_yO_zSi$) material is the commensurate flux of material that can control the buildup of polymer on the sidewall or vary its thickness within a feature (O, F, HF, H).

21.3.5.4 Post-Etch Metallization Pre-clean

While outside the scope of plasma etch, the etch process, followed by post-etch treatment to remove fluorocarbon residues from the dielectric surfaces, inevitably sees a copper surface exposed to oxygen. As Kojima et al.[93] point out, CuO, Cu_2O, and other oxides can form with oxygen exposure. Radical rich oxygen exposure (for example in an ashing process) tends to lead to hard-to-remove Cu_2O films whereas oxygen in the presence of ion bombardment (in an RIE plasma) tends to produce thinner smoother CuO films. These relatively tough surfaces have historically been removed through physical sputtering in a process known as "pre-clean." The trouble is that the sputtering process removes both field material and dielectric within a feature. It may redistribute unwanted material to sidewalls as well as compromise line to line electrical integrity. Advances such as reactive pre-cleans employing reactive gases to eliminate residue while maintaining etched structure integrity have substantially increased process integration reliability.

21.4 The Next Generation—45–32 nm Technology Nodes

As progress continues to be made to meet the ITRS requirements for the 90–65 nm technology nodes, significant technology challenges must be faced to continue this device performance improvement pace into the 45–32 nm technology nodes.[1] The 17% per year transistor delay performance improvement roadmap target will not only require new dielectric and gate electrode materials; but very likely require implementation of non-classical CMOS structures.[94–96] These will probably be non-classical multi-gate MOS field effect transistor (MOSFET)devices such as fin-body field-effect transistor (FinFET).[97–99] Similarly, circuit delay performance is ever increasing being impacted by the metal interconnect R/C line loss delays, so new ultra low-κ (ULK) dielectric materials may be introduced for the BEOL and possibly new structures such as an air gap integration may be needed to meet the roadmap targets.[100,101] These, along with the continual CD scaling, offer new challenges to etch technology. For example, new metal gate materials will likely be more difficult to etch while the selectivity requirements will increase as the film thicknesses decrease.[102] In addition, feature aspect ratios are expected to increase to maintain the same resistivity or to fabricate the novel structures. Some of the challenges to plasma etch technology related to high performance CMOS applications in the 45 nm node and beyond are now presented.

21.4.1 Patterning Related Challenges

21.4.1.1 Extension of Optical Lithography Roadmap with Immersion

The pending introduction of immersion optical lithography should significantly extend the expected use of 193 nm optical technology into the 45 nm and possibility into the 32 nm technology node with further improved higher numerical aperture optical lens systems and use of novel patterning schemes.[103] This

means that a completely new lithography technology such as EUV will be delayed.[104] For the etch technologist, this also means that the current chemically amplified resist will probably continue to be used, reducing the impact on the current equipment and process as the feature sizes are reduced. Limitations to processes introduced by LER, selectivity requirements, endpoint detection, etch rate controls as the feature dimensions and film thicknesses are reduced will be more important than they are now, compensating for the benefit of dealing with the same photoresist materials.[105] Even with the improved performance of 193 nm immersion technology, the feature size reduction may not be fully realized with single layer resist patterning and etch because of the difficult-to-etch materials and reduced resist selectivity.[106] Therefore, novel patterning approaches may also be needed to improve the ultimate (final) feature dimensions and control. For example, the photoresist trim technique currently used to reduce the minimum line width is limited by the 45 nm scaling spacing requirement and the increase in the etch selectivity requirement because of the loss of resist thickness. Multi-layer resists techniques briefly touched on in preceding sections have been shown to address a number of the process interactions to improve process capabilities and may further extend optical lithography.

21.4.1.2 Advanced Patterning Schemes and Possible Etch Interactions

The implementation of advanced patterning schemes have offered the opportunity to both extend the imaging capability by reducing feature sizing independent of the imaging process such as the improvements seen with the use of photoresist trim processing. Gate pattern photoresist trimming, discussed in Section 21.3.2.1, was introduced to allow the patterning of smaller features than can be resolved by lithography alone. Multiple layer patterning sequences such as silicon containing bilayer resists and hard mask where the multiple layers or steps will allow additional control or adjustment of the final pattern dimensions may be of greater importance. Primarily, the use of multiple layers allows the separation of the process interactions of the pattern generation and etch processes. For example, the use of bilayer or multi-layer imaging can be tailored to use very thin imaging photoresist that offers improved lithography resolution offering a tempting decoupling of etch and lithography processes.[107] Photoresists with increased sensitivity allow reduced exposure times that would otherwise not have the needed selectivity for the etch process. Likewise, the underlying etch resistant or hard mask layer can be dry developed or etched in less aggressive or more selective chemistry compared to the process required to etch the final feature such as metal gate electrode or thick oxide films. This separation of the processing interactions, less critical for 65 nm, offers multiple possibilities to optimize the patterning processes and to extend optical lithography several technology nodes.

Even with the addition of a separate masking layer etch or dry development step, there can be significant cost reduction by extending the life of previous generation process equipment. For example, using 248 nm bilayer resist with silicon containing image layer has been shown to have equivalent 90 nm contact yields compared to 193 nm single layer resist processing.[108] A bilayer resist process is also being used for advanced gate and STI patterning development. Continued development of multilayer resist and masking schemes should help extend 193 nm lithography technology capabilities as it has for instances of 248 nm technology. Recent developments include use of CVD carbon hard mask and CVD inorganic layers that double as antireflective layer and hard mask.[109] These offer the advantage of allowing the use of very thin photoresist imaging layer thus allowing a larger process capability to print smaller features for a given optical wavelength.

Electron beam (e-beam) direct write exposure, although probably not the lithography choice for volume manufacturing, could play an increasing role to be able to stay on the technology roadmap. E-beam exposure should be capable of imaging down to the 22 nm node and allow the parallel development of the other unit processes, device structures and product integration without the delays that may occur in the development of next level of optical lithography. In the past for example, delays in the delivery of 193 nm steppers added 6–12 months delays to the development timeline of the 120–90 nm nodes. There are increasing efforts to make e-beam system more cost effective, primarily by increasing throughput so that it may become a viable alternative where maximum resolution or minimum feature sizes are required and the optical solution will not meet these requirements. There is a significant time and cost impact of using direct write technique since the cost and time to product new optical reticles is

skyrocketing. This may reduce the dependence on the early development of optical solutions before introduction of the next technology nodes for product design and limited pilot production. For the etch technologies the main impact would be additional development due to the differences in the resist used for e-beam and optical exposure.

21.4.1.3 Aspect Ratio: Fundamental Concept

As the technology has continued to scale to smaller and smaller feature sizes, the concept of scaling the aspect ratio has become more critical in meeting the next technology node performance requirements. Whether this is a conductor feature where there is a requirement to reduce the size to shrink the circuit size and at the same time reduce the line resistance to improve circuit speed performance, there is usually a limiting feature aspect ratio where a physical constraint or other process requirement that will restrict the feature aspect ratio. These limitations involve many of the wafer processing steps such as lithography, chemical vapor or physical deposition, plating, and of course etch. For example, etching a thin layer that is approximately the same thickness as the feature dimension or an aspect ratio of 1:1 is much simpler than etching a very thick film at the same dimension. With the deeper feature, the selectivity to resist becomes more critical, etch lag may become a limitation, and feature sidewall slope may be too severe.

In creating manufacturing integrations to meet the design requirements, there have been some general rules of thumb applied that offer some qualitative evaluation to compare the technical difficulties expected when implementing a particular process or structure. A general rule is that the level of difficulty—risks factors, costs, etc., increases rapidly as the feature aspect ratio increases. Therefore, it is much easier to reduce the pattern CD if the photoresist thickness is reduced at the same rate; however, this is often not the situation. It is expected that this aspect ratio conflict between performance and process complexity will only increase with the scaling to 45 nm and below.

As illustrated in Figure 21.27, the ITRS Roadmap projects that the aspect ratio of the gate electrode may increase by almost two times from 3.2 at 90 nm to over 5.5 at the 22 nm technology nodes. So not

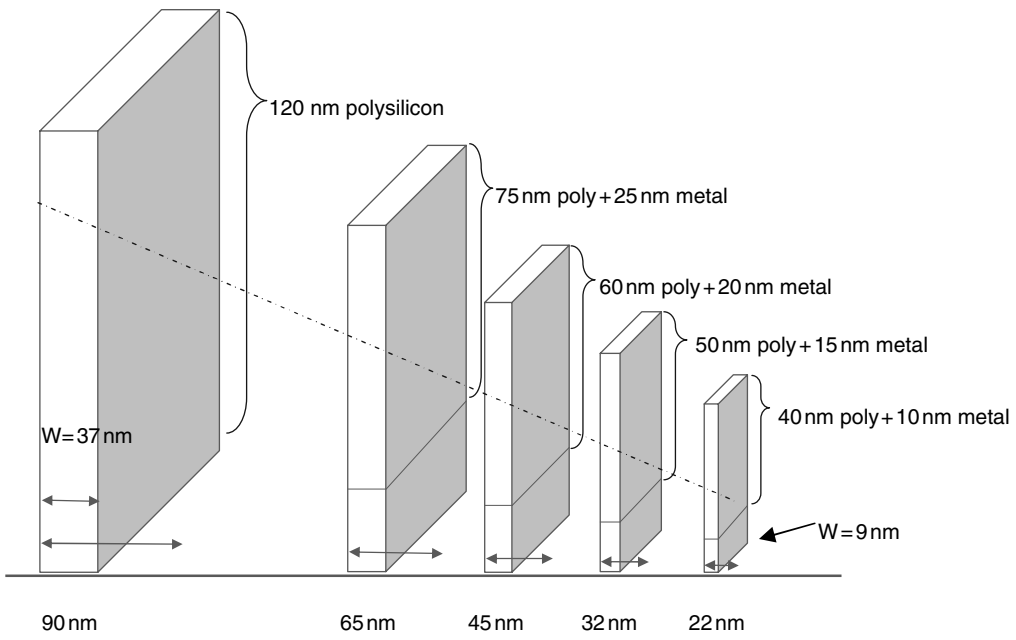

120 nm polysilicon
75 nm poly + 25 nm metal
60 nm poly + 20 nm metal
50 nm poly + 15 nm metal
40 nm poly + 10 nm metal
W = 37 nm
W = 9 nm
90 nm
65 nm
45 nm
32 nm
22 nm

FIGURE 21.27 Illustration of the impact of International Technology Roadmap for Semiconductors (ITRS) gate electrode scaling on the feature aspect ratio. The dashed line represents the 120 nm aspect ratio.

only does the CD scale down, the etch depth ratio will probably increase while the feature profile must be maintained for much more aggressive etch such as required for metallic gate materials.

Similarly, Figure 21.28 illustrates an even more difficult aspect ratio problem for contact or via features. Contact often presents an even more difficult challenge with increased aspect ratio since the first dielectric layer cannot be reduced due to minimum film thickness to provide isolation and protection. As seen above, the gate stack aspect ratio may likely increase, which will limit any reduction in contact aspect ratio as well.

Increasing the aspect ratio can quickly approach physical limitations that have not even been previously considered. Observations have shown the low temperature mechanical properties for photoresist deteriorates from exposure to even ambient moisture at high aspect ratio resulting in pattern collapse when the aspect ratio starts to exceed 4 or 5. These imply a number of other practical limitations such as photoresist distortion or deposition step coverage that may be not considered especially when defining only 2D circuit criteria.

In general, the aspect ratio may be related to a hypothetical process difficulty index (risk, cost, etc.). So in general terms when going to the next technology node, processing a feature aspect ratio of 1:1 may be considered to be trivial where the standard process capability for current technology can be reused. As the ratio increases to 1.5–2.5:1, the current process will likely require some additional process characterization, improvement or additional process control. Further increasing to the 2.5–3.5:1 range usually will require extensive process characterization or development, and often requires new monitoring techniques. And for aspect ratios of greater than 3.5:1, the next technology will often require new

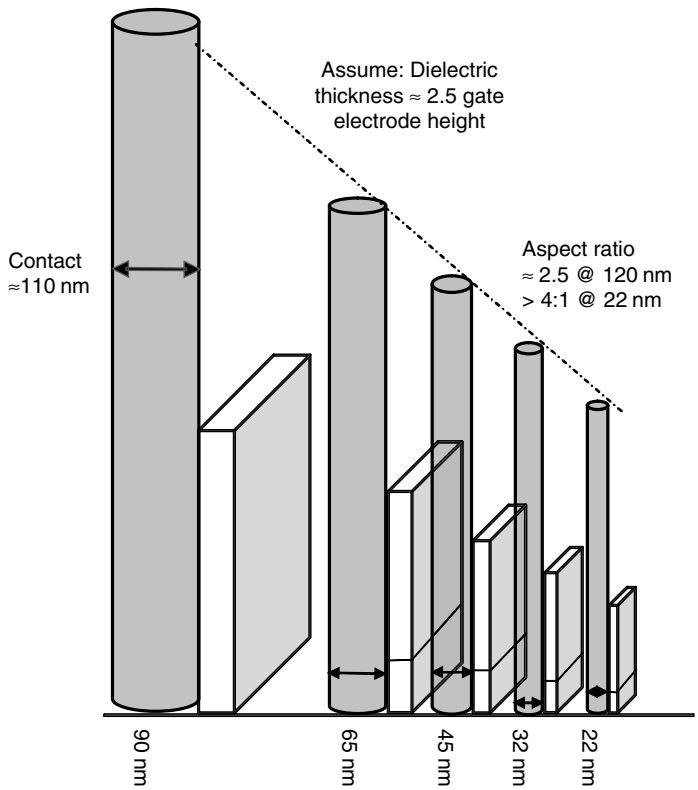


FIGURE 21.28 Illustration of International Technology Roadmap for Semiconductors etched contact feature scaling roadmap with the impact on the feature aspect ratio compared to the gate electrode. The dashed line represents the contact aspect ratio referenced to the gate electrode height.

process tools or new technologies for patterning, CVD or etch. So the same or preferably a reduction in feature aspect ratio can significantly reduce the risk and complexity for the introduction of the next technology node.

21.4.1.4 Imaging Resolution, Control and Line Edge Roughness

Assuming that immersion lithography in combination with alternative patterning techniques will be able to achieve the resolution requirements, then from a more practical standpoint, LER as defined in the final feature, may be the most critical technology challenge at the 45–32 nm technology nodes.[110] "Roughness" has not improved from the 90 to 65 nm nodes and in most cases it has worsened as device features have been scaled, the impact continues to become more critical. While a known phenomena at 90 and 65 nm, LER effects in device performance and reliability will become critical at the 45–32 nm technology nodes. A number of investigators have modeled and verified the electrical effects of the LER for various devices and have shown that it is not be significant effect compared to other process variations until 45 nm and below.[110–112] For example, a line CD variation of 8 nm (LER of 4 nm on each side) at 100 nm has very little effect on the transistor performance but at 30–20 nm it becomes the dominant variable. As for current and previous technology nodes discussed also in Section 21.3, phenomena like templating make sequences of etches, their interaction with photoresists in process integrations as important as single etch processes in role roughness plays in CD control. With respect to aspect ratio scaling, studies show that LER can be a function of photoresist thickness where thinner photoresist results in less LER—another case for the use of thin photoresist.

21.4.2 Device Roadmap Etch Challenges for 45–32 nm Technology Nodes

21.4.2.1 New Materials

Even though the transistor technology is discussed in other chapters of the handbook, it is probably worth a short review of some critical issues that may effect etch process development. Another approach to the device performance and critical scaling requirements is the introduction of novel device structures. These structures may compensate for material and process limitations such as patterning while meeting the device requirements. Novel transistor structures may likely dominate the scaling challenge due to cost and timing of the next generation lithography. As with alternative masking schemes, alternative device structures may provide a more timely and lower risk approach to meeting the device performance scaling challenges compared to the introduction of new materials into the classical planar MOSFET. These structures such as the FinFET transistor can be fabricated with existing materials—silicon and oxynitride dielectrics. Refractory metals and metal oxide dielectrics are more difficult to process, are expensive and pose cross-contamination risks.

At the 45 nm technology node, the device low power performance requirement will exceed the likely capability of planar device scaling with thermal oxide or nitride where the standby gate leakage may actually exceed the device on (ion) current. The reliability and the short channel effects may also limit the operating voltage requirements. To meet these requirements, the use of high-κ gate dielectric materials will be required. These offer new challenges to etch since these will most likely be a refractory metal oxide or silicate. Since pinning effects significantly limit the polysilicon gate material, the use of metallic gate may also be required. There may well be a metallic gate electrode etch that has to stop on a metallic oxide and both layers may be very difficult to etch.[113,114] High-κ material etch process chemistries are described by Sha and Chang.[115]

21.4.2.2 Novel Transistor Structures

There are a number of alternative or novel transistors being investigated including a number starting with the classical planar structure. For example, a fully silicide integrated gate electrode (FUSI) where the gate is formed in polysilicon and than a metal is deposited over the top of the gate. Then the gate is completely converted into silicide thus eliminating the need to etch a refractory metal. Improved device performance can be also be achieved by use of ultra-thin silicon active region operating in the fully depletion MOSFET

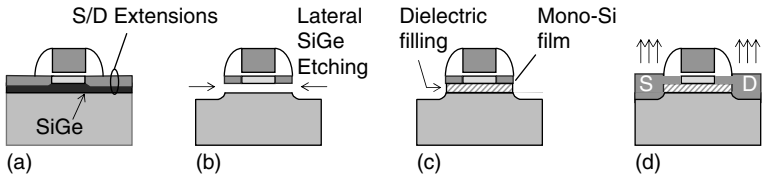


FIGURE 21.29 Illustration of the basic silicon-on-nothing process: (a) Selective SiGe/Si epitaxy and integration of the gate stack, (b) source drain (SD) areas etch and selective removal of the SiGe layer, (c) tunnel filling with dielectric layers, and (d) selective epitaxy of the SD areas to contact the Si-channel.

mode such as can be fabricated thin silicon on insulator (SOI). This approach requires only one mid-band gap metal gate material.

Also, a similar novel structure can be build using a sacrificial layer that can be replaced with a thinner dielectric isolation layer. The silicon-on-nothing (SON) transistor structure offers the possible advantage of using epitaxy on a bulk substrate that has more planar silicon integration; so most of the etch processes are consistent with planar CMOS.[116] However, to form the dielectric isolation under the active region, a sacrificial layer must be removed and a dielectric deposited under the active region, as illustrated in Figure 21.29. This is accomplished by the use of an high selective isotropic etch of a sacrificial layer of epitaxy SiGe under the thin active silicon layer. Since this is a lateral etch, it must be very isotropic and have very high selectivity to the very thin silicon layer. As similar novel structures are devised, new requirements may need this type of isotropic removal of thin sacrificial layers.

21.4.2.3 Ultra Thin Silicon Structures

One planar structure device that can address a number of the electrical scaling issues is the fully depleted transistor on ultra-thin silicon layer on insulator (FDSOI) planar structure.[99,100] Other than the aggressive gate CD scaling, the major etch issue with this type of structure is the use of very thin active SOI. The active silicon layer under the gate structure can be on the order of 100 Å, making the etch selectivity even that more critical. Since there must be a selective epitaxy layer deposited to reduce the source/drain resistance, the surface must have almost no damage or residue since there is very little material to be able to remove any damage or residue.

Multi-gate devices have been showed to have tremendous device improvement without the dependence on extremely effective oxide thickness scaling. But the fabrication of such double gate planar structures is difficult with the close gate alignment, contacting the bottom gate, multiple-layer processing, etc. Several self-aligned gate structures such as the gate all around reduce some of the alignment issues but involve very complex process sequences.

Multi-gate device (Figure 21.30)[117] are also interesting since the higher gate to channel ratio reduces the requirement on the oxide thickness and the narrow channel for speed performance. The device performance of multi-gate CMOS transistors has been shown by a number of investigators.[118] However, most of the original schemes when applied to the traditional planar device structures are very cumbersome to fabricate such as using a buried gate electrode. However, such a device that may be simpler to fabricate is a gate-all-around version of the SON transistor since it is a planar structure and the deposition of the gate electrode can occur after the sacrificial layer is removed under the active region. This also forms a self-aligned gate electrode.

To address these device issues, the vertical transistor or FinFET was developed since the active region is a very thin vertical element.[100,119–121] Even though this structure involves high aspect ratio features and topography, it utilizes standard thickness SOI substrates and oxynitride gate dielectric used for 65 nm high performance devices and has been shown to be extendable to possibly the 22 nm technology node.

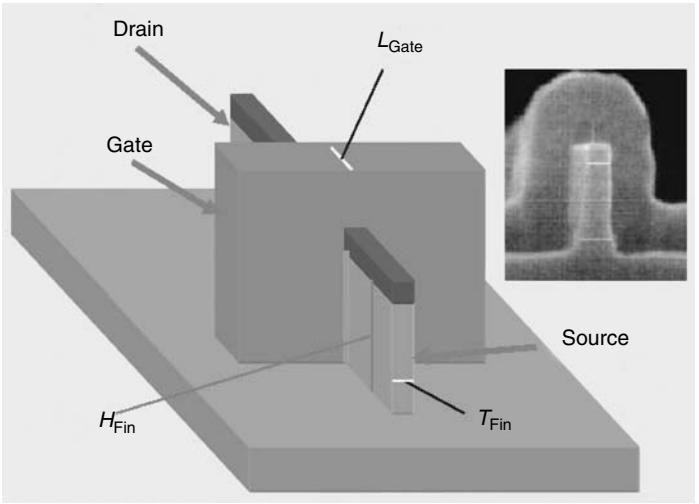


FIGURE 21.30 Illustration of the FinFET transistor. (Reprinted from Nowak, E. J., Aller, I., Ludwig, T., Kim, K., Joshi, R. V., Chuang, C.-T., Bernstein, K., and Puri, R., *IEEE Circuits Dev.*, 20, 20, 2004. With permission. Copyright 2004 IEEE.)

21.4.3 Front End of Line Etch Processes

21.4.3.1 Gate Etch

The application of metallic gate electrode materials with high-κ metal oxide or silicate dielectrics is very likely the dominant technology for the classical planar complementary metal-oxide-semiconductor field-effect transistor (CMOSFET) device structures going into the 45–32 nm technology nodes (ITRS roadmap). These materials are needed to meet the device speed and dielectric leakage requirements. Most of the refractory metals, silicide, and metal oxide materials identified as possible candidates are more difficult to etch than polysilicon since these typically have less non-volatile etch by-products and the reduced effectiveness of a hard mask at these dimensions. The use of selective epitaxy in active regions may increase the requirement for less etch residue and mechanical damage.

21.4.3.2 Dual Metal Gate

Maximization of CMOS performance for the classical planar device structures will probably also require dual metal electrode materials with matched work functions for both n- and *p*-type devices (Figure 21.31). Not only does this mean additional etch processes for quite different materials, but depending on the integration, both materials may have to be etched as a stacked layer. This also places additional selectivity requirements on the opposite doped type regions. To attempt to simply the integration, one metal gate material is deposited on the high-κ dielectric and then removed where the other type devices will be fabricated. Then, the next gate electrode is deposited usually followed by polysilicon deposition. This means that the first etch must not etch or damage the high-κ dielectric that will have detrimental effects on the devices made over this dielectric. The gate patterning must etch both metal layers on some of the die and only the second metal over the remaining part (see Figure 21.31). Again, this places significant selectivity challenges for the gate etch since the area with one metal layer will be exposed to additional etch of the high-κ dielectric.

21.4.3.3 High-κ Dielectric Etch

Following the gate etch and usually the spacer formation, the high-κ dielectric must be removed. Unlike the classical silicon oxides, these materials do not etch in HF solutions and often depends on the amount

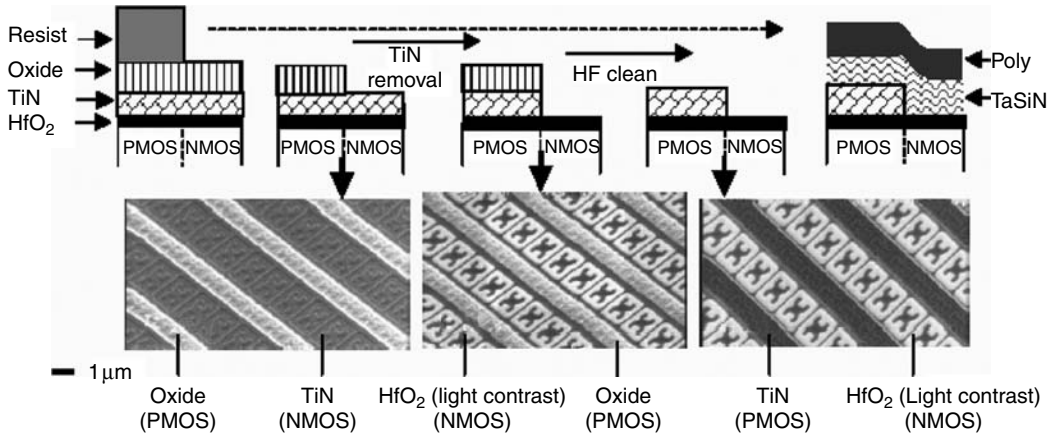


FIGURE 21.31 Schematic of a dual-metal gate process flow with corresponding scanning electron micrographs (SEM) images of a static random access memory (SRAM) bit cell array showing the different layers for the N-channel metal-oxide semiconductor (transistor) and P-channel metal-oxide semiconductor (transistor) regions. (From Samavedam, S., La, L., Smith, J., Dakshina-Murphy, S., Luckoski, E., Schaeffer, J., Zalava, M. et al., *Symposium on VLSI Technology*, 24, 2002.)

of damage during the gate etch or other process conditions such as thermal treatments. Similar to the metal gate materials, these are also difficult to dry etch as they can have some morphology and have metallic component that is less volatile such as hafnium or zirconium, and as a result the selectivity to silicon substrate is typically not as good. Unfortunately, due the scaling issues and fabrication of the transistors on ultra thin silicon regions the required selectivity is significantly higher. Considerable effort is underway to identify possible dry etch solution but as no totally satisfactory solutions have been developed, a different approach such as atomic layer etching (ALE) may be required.

However, it is difficult to focus on all the possible etch schemes since there is not a clear technology direction that will be taken by the industry. But there are some clear challenge differences for some of the likely alternative options, and these should be considered in the etch technology development and ultimate capabilities or these will define the next etch technology. For example, the planar, ultra thin body transistor structure or SON requires the removal of a very thin sacrificial layer under the very thin active silicon region[116] (Figure 21.29). This etch by definition must be highly isotropic. Wet chemical etching has been used but this is very limited and the mechanical stresses on these extremely thin structures may be unacceptable except for investigation and feasibility tests. A chemical downstream etch has been shown to give highly isotropic etch using fluorine chemistry such as CF_4.[122] However, this is a timed etch since it does have an effective end point capability, and there is a loss of selectivity when the SiGe is completely removed (endpoint). An alternative process using an ICP chamber operated in a remote low power plasma mode with no bias power.[123] This has been shown to the same or higher selectivity with improved uniformity and the capability to monitor the plasma using optical emission.

Probably the leading candidate for the next generation novel structure is the vertical dual-gate on ultra-thin silicon or the FinFET structure (Figure 21.30). There are a number of different multi-gate, dual gate, or tri-gate variations that have been proposed, and each offer some slightly different etch challenge. The most basic structure will probably be a simple vertical fin formed in a SOI substrate with initially classical doped polysilicon gate electrode on silicon oxynitride dielectric. The silicon etch uses typical gate chemistries except that it is a very high aspect ratio, 10:1 and greater, depending on the device performance requirements. This has many of the concerns expressed about the difficulties of processing high aspect ratio features.

21.4.3.4 Spacer Etch

Another process that has been impacted is the gate spacer etch. Although very similar to previous technology nodes, this process is also affected by the feature scaling. This means that not only will the spacer be narrower with a higher aspect ratio, but also the uniformity requirements will increase to reduce the variation on the device performance. Similar to the gate etch, it is more difficult to etch with very high selectivity to be able to stop on the high-κ gate dielectric compared to silicon oxynitride films. Also, on novel structures such as the FUSI or FinFET, there may be multi-type film stacks such as metal gates or even more extremely high feature aspect ratios.

Of course, there is still the inevitable device scaling, which puts more requirements on the hard mask and photoresist trim process capabilities. As mentioned previously, a key to improved patterning process capability will be the use of multi layer patterning especially with an organic hard mask layer. This will probably be more compatible to the new gate materials and the lower thermal budget that may be required. Also, the multiple layers allow the opportunity to optimize the feature sizing methods such as trimming.

21.4.3.5 STI or Isolation

For those devices that may still use bulk silicon, there is even more CD control and probably the same LER requirements as the transistor gate. Improved transistor isolation will mean even higher aspect ratios and profile controls since the use of low-κ dielectrics maybe needed to improve cross talk similar to the metal interconnects. These typicallly have the same feature gap fill issues as the current TEOS CVD films but be more difficult to implement for the planarization and oxide recess through the following etch and clean steps. Low-κ dielectric even Fluorinated TEOS (FTEOS) introduces new concerns about other chemical effects or "contamination" of the active region. Recent device performance has been seen with the use of a stressor layer in the bottom of the STI regions. This again adds complexity and control requires for the etch and module integration.

There will likely be a number of silicon etch processes required for the various electrical isolation steps such as for FDSOI transistor. The trend to more SOI structures, even to ultra-thin silicon structures offer actually less challenges for etch since the etch needs only to pattern very thin silicon layers over BOX. In fact this is a significant factor in selection of different device integration strategies.

21.4.3.6 Contact Feature Etch Challenges

Contact is always a very critical closed feature that has to be fabricated since it has the minimum sizing to allow the highest possible circuit density. And usually it is a very high aspect ratio feature since as mentioned before, the gate electrode stack and dielectric capping layers are not scaling with the technology. Similar to the current technology nodes, the requirement for contact and interconnect become more difficult with scaling but contact is more difficult to image with extremely low defectivity or missing patterns. A relatively thick photoresist is needed for the high anti-reflection (coating) (AR) oxide etch process, so a high photoresist exposure dose is required and, along with high pattern density image proximity effects, results in a small process window. These factors have prompted the increased use of silicon containing bilayer or multi-layer resist strategies at the 65 nm node and will probably become required at 45 nm and below. Also the roughness (similar to LER) that occurs during patterning is also a significant problem at contact since this marginality in the process is indicative of a high defective process and can also impact contact resistance and reliability. So the lithography challenge is some ways except for CD uniformity is more difficult than gate since there is not a post-develop "trick" that has been shown effective in reducing the printed image sizing. Some techniques such as material "repair" or treatment have been shown to improve the resist selectivity after develop but these do little to reduce the sizing. There have been a number of attempts to increase the post-developed contact sizing by the addition of a polymer (resist) that reacts with the contact sidewall similar to the resist poisoning phenomenon. After a second develop step, there is a layer remaining on the sidewalls of the original photoresist that effectively reduces the contact pattern CDs.[124]

21.4.4 Back End of Line Etch Processing and Ultra Low-κ Dielectrics

As device scaling has meant the increase in device speeds, the BEOL has not been a significant limitation to circuit performance until recently. Significant performance effects have been seen at 90 and 65 nm high performance circuits due to the BEOL resistivity and capacitance increasing line capacitance coupling. At 45 nm some have suggested that this may actually be a more significant limitation than the transistor performance. The trend in the industry to compensate this effect has been to introduce low-κ dielectric materials. But this is proving to be more difficult than expected and the copper line resistivity has been shown to also be a critical issue. The copper line resistivity is increasing more than predicted for a given scaling factor. There are several effects such as barrier metal thickness, as well as significant grain boundary and surface effects at these dimensions. So the BEOL metallization may become the new limitation for continuing on Moore's Law scaling roadmap.[125]

This scaling requirement for 45 nm requires smaller CD but, for lower copper line resistivity, this means deeper metal trench structures or higher aspect ratios. But as the line depth increase so does the line capacitance losses, which requires the use of even lower-κ effective dielectric structures or integrations and to maintain this during processing. It also could result in the use of a thicker barrier to maintain step coverage or mandate the use of a different process to improve step coverage such as atomic layer deposition (ALD). As expected etch can have a direct impact on these parameters.

21.4.4.1 Challenge of Etching Ultra Low-κ Dielectrics

The porous or porogen containing ULK materials under investigation for the 45 nm BEOL generally etch much faster than previous inter-layer dielectrics such as TEOS or carbon doped oxide (SiCOH). However, there a number of issues with these materials that impact our ability to etch. For example, these films' dielectric properties are very sensitive to subsequent processing such as CMP, cleaning and photoresist removal. A more dense oxide cap layer is therefore used to protect the bulk film. But as the structures are etched, the exposed sidewalls are subject to the damage. To prevent this sidewall damage, various modifications have been made to the processing such as the ashing and etching chemistries.

Another approach is to help minimize the exposure of the low-κ material. A dual hard mask integration approach helps since the trench pattern is etched in the top hard mask layer and the resist can be stripped without exposing the low-κ to the plasma, usually a standard oxygen process. In this trench first sequence, the via pattern is then applied and etched into the second hard mask layer and through the low-κ dielectric, stopping on the etch stop/diffusion barrier at the bottom. The via layer resist is removed while only the less critical via sidewalls are exposed. The trench features can now be etched through the top hard mask layer and the etch stop layer (ESL) opened to the copper. This improves feature alignment and reduces the exposure of the metal sidewalls to oxidizing chemistries that can significantly impact the dielectric properties.

In addition, considerable investigation has been done into alternative photoresist removal processing and chemistries.[126,127] Most of these rely on reducing chemistries such as NH_3 or N_2/H_2, or on less oxidizing mixtures such as CO or CO_2/O_2. Photoresist removal can also can be done in situ or immediately following the dielectric etch in the same chamber, but there can be significant chamber wall effects depending on the etch chemistries and reactor design. If reactive polymers are deposited on the sidewalls, then these can be released during the photoresist removal step increasing the etch or damage to the low-κ material sidewalls.

21.4.4.2 New Porous ULK Dielectrics

To reduce the emissivity of the dielectric, the pore density and size must be increase and this leads to obvious impact on the mechanical properties. This worsens the etch sidewall damage since more of the layer becomes carbon-like or porous. To help alleviate some of the mechanical and process issues with the porous dielectric films, a different low-κ material is being developed where the porosity in not introduced until after the etch and metal deposition process steps. This is referred to as a "solid first" approach.[128]

The as deposited film contains clusters of a polymeric material that decompose at low temperature than the bulk film. So after CMP, metallization and dielectric deposition processing such as the cap layer or several metal layers, the layer is heated to above this temperature and the porogen decomposes leaving voids in the matrix dielectric layer. This reduces much of the mechanical issues since the porogen containing film is denser and has higher modulus. Also, the lack of pores reduces the diffusion of reactant species that can remove carbon within the layer or damage the layer during etch and resist removal. The dielectric etch performance is also improved since the film density results in better etch profile control. This approach has a number of processing advantages, but there are still many questions concerning the reliability due to high film stress that results from the significant shrinkage of the film during the porogen removal process.

21.4.4.3 Metal Line Resistivity

A significant line resistivity effect is due to the thickness of the diffusion barriers such as tantalum, tantalum nitride or TiN.[129] Even though this is primarily a deposition process limitation, the side wall roughness and more importantly the profile shape following etch can be significant in the ability of given deposition process to provide continuous uniform barrier coverage. As mentioned before, there is often an oxide or cap layer used to protect the ULK material. With the high etch rates for the low-κ dielectric and the possible residual chamber effects a significant undercut or notched profile can occur during etching and resist strip. This can be very problematic for deposition even ALD.

Again LER and sidewall texture can contribute to part of the increasing line resistivity by requiring thicker barrier thicknesses to have the required reliability. Porosity of the ULK dielectric also contributes to this issue and some have even proposed a separate deposition to "seal" the pores to improve barrier coverage and limit metal diffusion. Lastly, the effects of surface scattering is only made worst for rough surfaces.

21.4.4.4 The Ultimate Low-κ: Air

As the challenges increase to find a material that meets the dielectric properties and that can be processed while maintaining the required effective κ value, many as the best approach consider a structure change. This is similar to high performance transistor where the integration may have to resort to using a novel structure instead of difference materials such as the structure illustrated in Figure 21.30. As previously mentioned, the BEOL interconnect may offer biggest challenges to meet the technology performance requirements at 45 nm and below. Even lower emissible dielectric material will probably be available but the integration of these into a more manufacturable process seems to offer un-surmountable difficulties.

For most of the proposed air gap integrations there is actually a significant reduction in the challenges to the etch processing since these usually would reuse previous generation materials such as TEOS, low-κ dielectrics or sacrificial organic layers (Figure 21.32). These have previously been extensively utilized in previous generations and the etch processing is well characterized and understood. In some proposals, organic layer can be used as the metal line layer such as the porogen material so the line etch is basically similar to the dry develop process used for similar multilayer resist processing. For most air gap structures, there remain several issues besides the inherent reliability of these structures including the etch and metallization process interactions of having unlanded vias on air spaces.

Although this may seem to be a significant issue, in fact in practice, these may not actually be as critical. First, most of these can be mitigated by layout, sizing and the use of via etch processing that provide taper or smaller bottom CD's. In most processing, the effects of unlanded via structures have been observed and these effects are more significant in the porous low-κ materials due to the higher etch rates. In misaligned test structures, deep very small voids are etched along the sides of metal lines, but in almost all analysis these do not form voiding of the subsequent metallization processing nor have via failures been reported for such features. In actual circuits layouts, there are other more significant effects of such misalignment such as resistance or shoring that require sufficient alignment and overlay control where

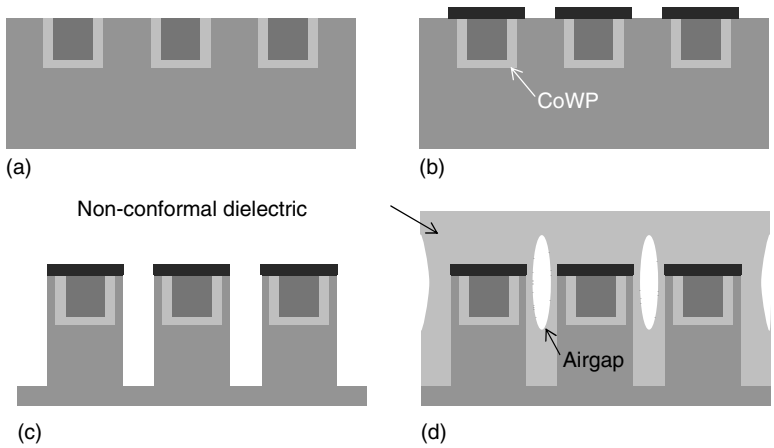


FIGURE 21.32 Schematic representation of air gap formation using non-conformal chemical vapor deposition (CVD) technique: (a) Once the metal layer is completed, (b) a self-aligned cap layer is deposited on the copper lines, (c) the dielectric is etched between the line (d) followed by a non-conformal CVD process to form air gap spaces.

this misalignment has not been a problem for full dielectric structures. But more data is needed for the air gap structure to determine the actual mechanisms and the possible impact on reliability and yield.

21.5 Nanotechnology—22 nm and Beyond

21.5.1 Device Roadmap Challenges

In the past several years, a number of investigators have predicted the end of the silicon era as we approached the threshold to nanotechnology. Even the definition of where nanotechnology begins is a topic within itself, but most agree that the 22 nm node is well at this technology. It is very risky to attempt to project the future, but that is more the case for semiconductor technology development trends, especially more than one generation out. If the 45–32 nm nodes will be the transitional phase with the introduction of some new materials and structures, there is little disagreement that the 22 nm node will be at the nanotechnology level and maybe at the limit of silicon technology.[94,110,130,132–134] Since every physical limit of the conventional technology will be pushed if not exceeded, new structures and materials will be needed at almost every element to be able to build such structures, but it appears that silicon will still be the fundamental base for the technology. The devices will have to be scaled truly to the atomic level as illustrated in Figure 21.33 and Figure 21.34, and so the process technology must also start to perform at this level.

A number of investigators have attempted to forecast the most likely path and, if we assume a normal 15 year Intellectual property (IP) development cycle,[135] these projections should start to be a bit more focused on the most likely scenarios. These forecasts usually involve the transition from classical planar CMOS on bulk silicon substrate to planar on ultra thin silicon active region such as FDSOI. This will probably quickly migrate or will be introduced with high-κ dielectrics and metal gates if the manufacturing issues are resolved in the next several years. The next performance boost will probably be similar devices but with multi-gates.[136] These seem to be consistent with the ITRS roadmap and could be introduced later in the 45 or 32 nm nodes.[1] The possible alternative to this migration is the FinFET structure as described earlier (Figure 21.33). It offers the advantages of both ultra-thin silicon and multi-gate performance. And because of the simpler technology, it can provide the needed performance but much sooner than the next level of scaled planar technology will likely be available. Nanowire transistor

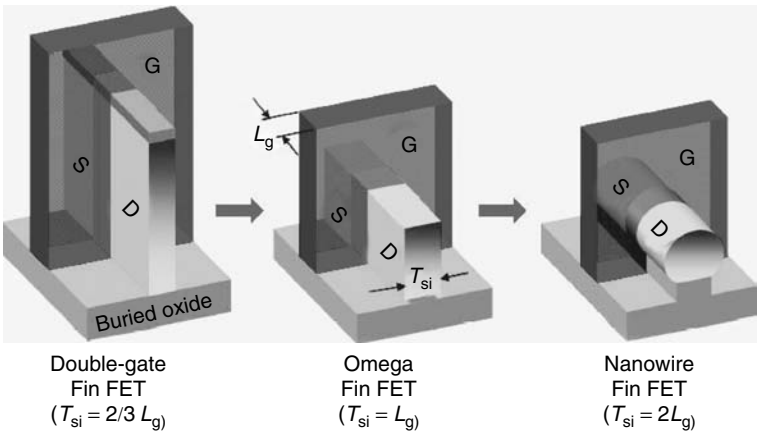


FIGURE 21.33 Evolution of FinFET device to nanowire structure. (Reprinted from Yang, F.-L., Lee, D.-H., Chen, H.-Y., Chang, C.-Y., Liu, S.-D., Huang, C.-C., Chung, T.-X. et al., *IEEE Symposium on VLSI Technology*, 196, 2004. With permission. Copyright 2004 IEEE.)

could then be the next evolutionary extension of the current FinFET structures and is projected to provide further enhanced device characteristics so that the extreme scaling (beyond the lithographic limit) may not be needed[137,139] (Figure 21.34).

Not only would this be a step function in the technology roadmap, but also most investigators now see this as the logical progression to continue the device scaling of a 45 nm FinFET that can be scaled to 22 nm nanowire transistors. The introduction of carbon nano-tube FET carbon-nanotube field-effect transistor (CNFET) could also use the same basic structure with the silicon nanowire active region being replaced with carbon tube, so yet further scaling to single electron quantum dot structures[134,140–142] (Figure 21.35).

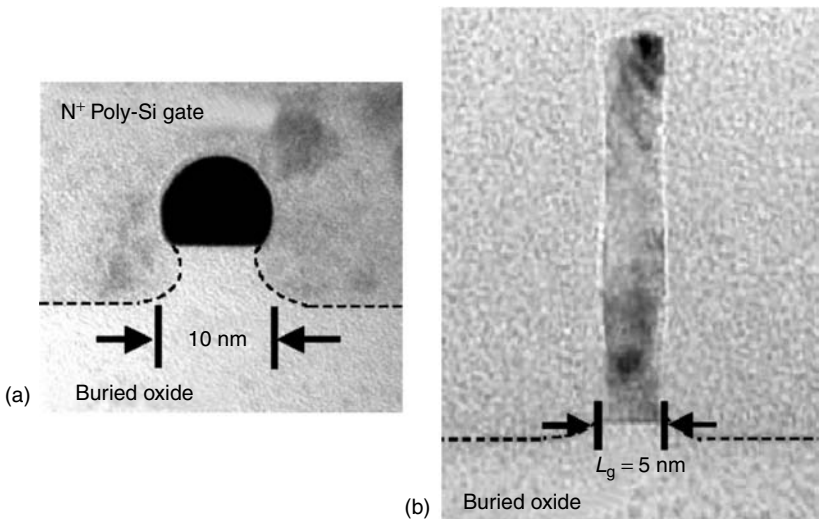


FIGURE 21.34 Transmission electron microscopy cross-sections of nanowire structure. (Reprinted from Yang, F.-L., Lee, D.-H., Chen, H.-Y., Chang, C.-Y., Liu, S.-D., Huang, C.-C., Chung, T.-X. et al., *IEEE Symposium on VLSI Technology*, 196, 2004. With permission. Copyright 2004 IEEE.)

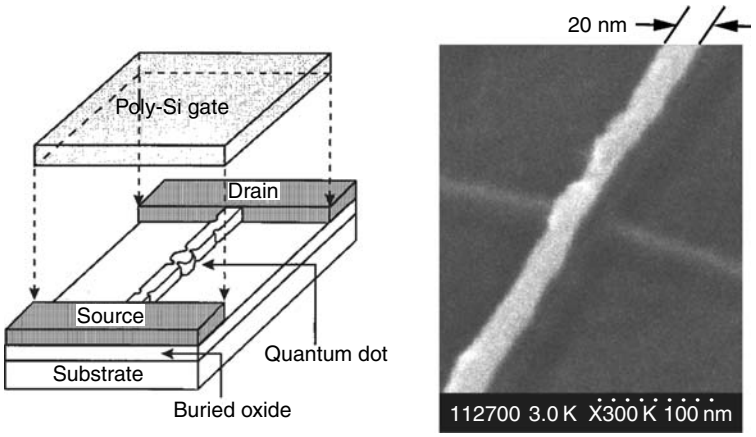


FIGURE 21.35 Extension of FinFET like structure to quantum dot device structure. (Reprinted from Zhuang, L., Guo, L., and Chou, S. Y., *Appl. Phys. Lett.*, 72, 1205, 1998. With permission. Copyright 1998, American Institute of Physics.)

If many of the FEOL roadmap limits seem to be have been pushed out, the critical BEOL interconnect is just now being addressed - increasing copper line resistivity.[143] Even at the 32 nm node, the increase in copper resistivity may actually limit circuit performance. There does not appear to be any obvious solution to the resistivity issue at 22 nm since ultra thin barrier, high aspect ratio, and design options will need to be pushed to their extreme to allow 32 nm to meet the technology requirements. The only possible migration will be in further reduction of the line-to-line capacitance. As previous discussed, low dielectric emissivity can significantly improve the resistance-capacitance product (RC) delay losses and air gap can approach κ values close to 1.0.[133] So as the resistivity continues to increase without much opportunity to reduce it, improving the interconnect losses at 22 nm must take advantage of the air gap structures. Besides the issues previously discussed for air gap structures, the requirement to attempt anything that will reduce resistance will likely drive the thickness of the lines up that will further strain the patterning requirements with increasing feature aspect ratios.

Extending the previous materials discussions, there is in some ways even less uncertainty about the alternative materials for the 22 nm technology than for 45–32 nm; because for now, the scaling can only be accomplished with a FinFET type structure for the basic CMOS transistor. The materials needed to meet the device requirements will probably be developed at 45–32 nm for the FDSOI and other planar devices. For example, the FinFET devices are capable of meeting the 45 nm targets with polysilicon and silicon oxynitride but by the 32–22 nm nodes, these too will probably require high-κ dielectrics and metal gate electrode materials. Since there are already feasibility demonstrations of device structures at the 22 nm scaling, there is a high probability that new technologies such as ALD can be scaled to this level.

21.5.2 Imaging Limitation and Etch Interactions

Beyond extending optical immersion lithography to the 32 nm node, an alternative approach for gate fabrication may be the use of a hard mask formed by a spacer process on a sacrificial material.[138] This has some definite limitations and requires double patterning—one for the minimum gate CD and one for the contacts and larger features, i.e., anything greater than the minimum. As in the current lithography development thrust, LER becomes even more challenging for lithography and etches with $\sim$9 nm physical gate length and total gate roughness of 1 nm.[1]

21.5.3 Extension of Existing (45–32 nm) Etch Technology Node

As with the 45–32 nm technologies, the critical etch challenges will probably be the feature aspect ratio and selectivity requirements to be able to meet the scaling roadmap targets. With appropriate masking technique resolution at 22 nm does not appear to present any significant roadblocks but still a number of challenges. Unfortunately, the aspect ratio become even more difficult with a number of critical steps approaching or exceeding 4:1. Beside the contact etch ratio that will probably exceed 4:1, the BEOL copper metal resistivity is increasing as the line become smaller and this will drive the integration to have deeper metal trench structures as well.[102]

Even though 22 nm will use material processing characterized for previous generations such as high-κ dielectrics and metal gates, the selectivity challenges are even more difficult since the layers become extremely thin. These will be even more difficult along with the probable use of 3D transistor structures may require a different approach—ALE. This is analogous to ALD now being developed or already introduced into standard CMOS technology. Atomic layer etching can be modified to provide etch conditions ranging from highly anisotropic to purely isotropic. For 3D structures such as the SON transistors previously discussed and nanowire structures, isotropic etch processes may be needed and these will require the same precision control of any anisotropic process[144] (Figure 21.36).

Primarily chemically driven selective etch processes may no longer provide the capability since in many cases the polymer level through which the etch chemistry reaction must proceed is thicker that the feature size. So like with ALD, the ALE reactant species must be adsorbed on the surface, and then energy is supplied to cause these molecules to react. So for each layer or cycle, the chemistry and reaction rate can be precisely controlled. Use of thermal energy such as with a heated substrate stage or blanket radiation such as in rapid thermal processor will result in isotropic etching. The use of a directional energy source such as ion bombardment, electron beam or collimate photon beam will result in very anisotropic process. Since many of the materials may have similar composition such as metal and their oxides, ALE does not need to have a chemical selectivity mechanism; but in principle, the desired number of atoms is removed per cycle.

Even in those etch processes that do require quite the same level of control selectivity or do not lack the chemical etch selectivity properties, the 22 nm technology will still require significant improvements in process control. As discussed in an open panel forum at the American Vacuum Society Symposium 2003, industry process technology leaders expressed the need for a paradigm shift in equipment and process control—from one of controlling equipment parameters such as power, pressure, temperature to one of being able to adjust ion energies, into densities, ion species, etc. One technique that offers most of these is neutral beam etching. Similar to ion implantation where the beam properties can be precisely

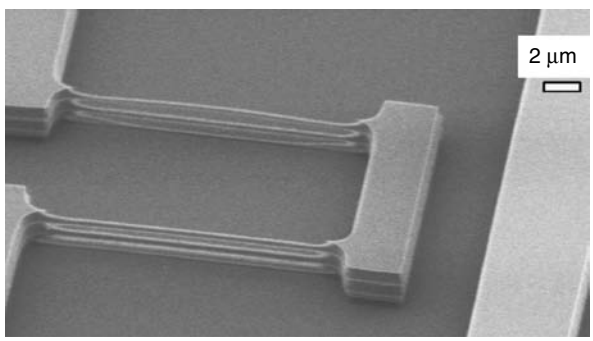


FIGURE 21.36 Example of a 3D nanowire structure fabricated using isotropic etch. (Reprinted from Milanovic, V. and Doherty, L., In *Proceedings of ASME International Mechanical Engineering Congress and Exposition* (*IMECE*), 33392, 2002. With permission. Copyright 2002, ASME.)

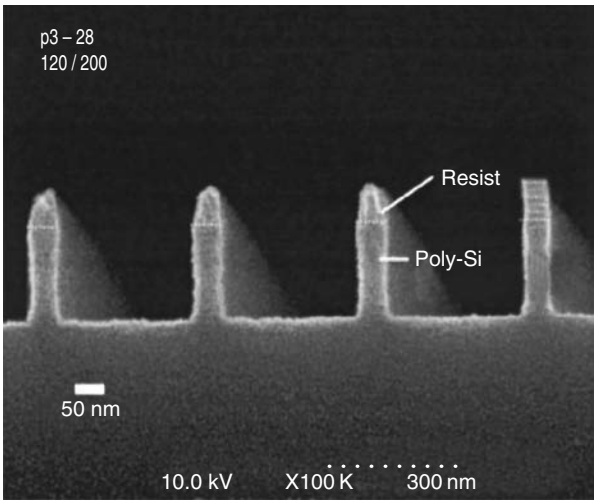


FIGURE 21.37 Polysilicon profiles using Cl_2/SF_6 neutral beam etch. (Reprinted from Noda, S., Nishimori, H., Ida, T., Arikado, T., Ichiki, K., Ozaki, T., and Samukawa, S., *J. Vac. Sci. Technol. A*, 22, 1507, 2004. With permission. Copyright 2004, American Institute of Physics.)

determined, so does the large area ion beam provide similar controls but for the entire substrate to be able to maintain low process times.

Although neutral ion beam may see limited introduction for 45–32 nm technologies, it is difficult to imagine that the current RIE technique will be extendable to the 22 nm level for most critical etch applications. S. Noda et al. have reported excellent results using neutral beam processing of polysilicon gate etch[145] (Figure 21.37). One such early application of neutral beam processing has been the demonstration of photoresist strip on ULK porous dielectrics.[146,147]

21.5.4 Process and Equipment Requirements

It has been shown that generally new process equipment is needed about every other generation.[131] This is true for deposition with the probable wide use of atomic level deposition at the 45–32 nm technologies for the deposition of high-κ gate dielectrics and copper BEOL diffusion barriers. This equipment scenario will probably be the case for lithography and etch at the 22 nm node. For etch, the need to have increased selectivity and atomic level control will drive the introduction of ALE. This equipment could be the same used for deposition today with slight optimization for the chemistry and reaction mechanisms. In addition, current etch tools could also be easily modified to include the capability to cycle gas injection and ion bombard for an ALE process while maintaining the flexibility to run more tradition RIE conditions. However due to the very low etch rates, these ALE processes will probably be limited to very thin films where there is not a good chemical selectivity mechanism.

For other processes with either every increasing aspect ratio such as contact etch and the thicker layer etching where ALE may not prove viable, the stringent process selectivity and feature definition requirements still must be achieved. These requirements will probably drive equipment designs to allow lower and selectable ion energies, and maybe the possibility of selecting the ion species and higher directionality. An example of a critical process is the FinFET fin and gate etches where the aspect ratio is increasing while the CD is rapidly scaling down. Also, the gate etch must be done over the topography of the fin structure as illustrated in Figure 21.38 and serves as another example of the process complexity tradeoff as a function of feature aspect ratio.[133] As mentioned already, neutral beam etch seem to offer

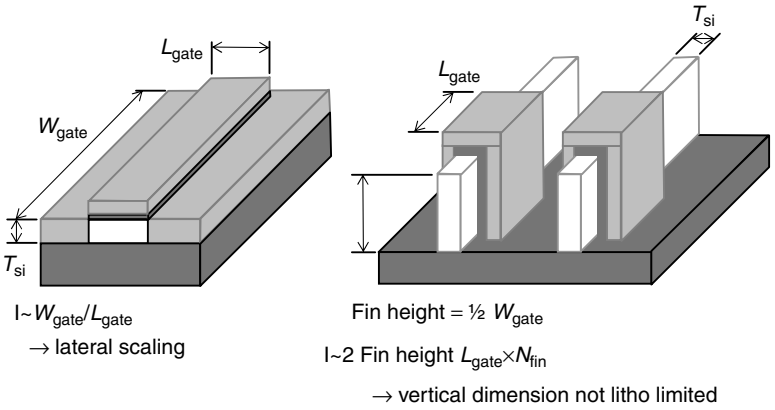


FIGURE 21.38 Comparison of planar device and FinFET aspect ratio and scaling requirements.

this level of ion level control. The basic design is simply an ICP source with an ion extraction grid or a high aspect ratio "neutralization" plate (Figure 21.39). There are number of ICP sources that can be readily modified to add this feature.

21.5.5 Challengers to Silicon beyond 22 nm

There is already intense research into the alternative materials for silicon; but of these, the carbon nanotubes is probably the most likely next material. There are a number of properties that exceed those of silicon such as resistivity, strength, etc., but the more important is that it appears to be compatible with

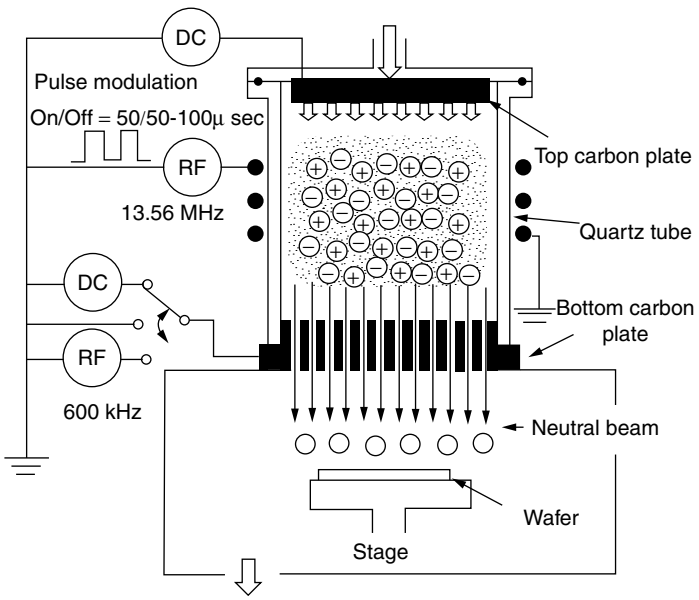


FIGURE 21.39 Schematic diagram of an experimental neutral beam etch chamber. (Reprinted from Noda, S., Nishimori, H., Ida, T., Arikado, T., Ichiki, K., Ozaki, T., and Samukawa, S., *J. Vac. Sci. Technol. A*, 22, 1507, 2004. With permission. Copyright 2004, American Institute of Physics.)

silicon processing. For example, it could be used in the standard process as the active channel region for the transistor where the source-drain and gate element can be fabricated in silicon.[148,149] Due to the low resistivity and mechanical stability, carbon nanotubes may also find applications in the BEOL such as via since these may facilitate the columnar growth.[150,151] However, there is almost as intense effort ongoing with silicon structures that could extend similar nano-column or tube structures in silicon, so only at time in the future will there be an end of the roadmap for silicon etching.[134,135,152]

21.6 Modeling of Plasma Etching Processes

Plasma etching processes are physically and chemically complex phenomena, and are often difficult to thoroughly characterize experimentally. Considerable attention has therefore been paid to computational modeling of plasmas and etching processes in the last 15 years. Some aspects of plasma processes can now be reliably analyzed using commercially available software, while other areas remain topics of intense research. With growing complexity of plasma processing applications, introduction of new materials at an unprecedented pace, and structure dimensions approaching nanometer scale, it is imperative that modeling will play a strong role in design of future plasma processing tools and processes. A brief review of plasma models is included in this section. More details can be obtained in the cited references. Attention here will focus only on computational models. The importance of analytical models (e.g., Refs. 2 and 153) can however not be deemphasized as they remain invaluable tools for plasma tool design and engineering analysis.

Computational plasma process models can generally be sub-divided into three categories: equipment models, feature scale models, and atomistic models. These models are often inter-coupled to analyze complicated problems but large disparity in time and spatial scales makes simultaneous simulation of all pertinent physical and chemical phenomena very challenging. Equipment models typically address gas flow in the plasma reactor, plasma generation, chemistry within the plasma, reactor electrodynamics, plasma interactions with driving circuits and sheath dynamics. These models address phenomena on relatively large spatial scales (cm) and moderate time scales (ns–ms). Feature scale models simulate etching and related surface phenomena within small structures. Analysis often relies on macro-variables (sticking coefficients, sputtering yields, etc.), to represent surface processes. Feature scale models address issues on small spatial scales (μm) and relatively long time scales (seconds). A new class of models utilizes molecular dynamics (MD) or ab-initio techniques to investigate etching relevant surface processes from first principles. These models address issues on very small spatial (nm) and small temporal (fs–ps) scales.

Kinetic, fluid and hybrid techniques were all explored in the early days of multi-dimensional plasma reactor modeling. Kinetic models include particle-in-cell models[154] which self-consistently track macro-particles and their interactions and models that attempt to directly solve the Boltzmann equation[155] to determine important plasma properties. As these techniques are computationally expensive and it becomes progressively difficult to represent the complexity of actual plasma processes, these techniques are primarily used now for research or specialty applications. Fluid[156] and hybrid[157] techniques however have been explored in considerable detail and several commercial software[158,159] based on these techniques are available. In fluid and hybrid plasma model, Maxwell equations are solved in conjunction with equations governing species mass, momentum, and energy balance to determine important plasma properties. Electrons generally drive etching relevant plasmas and have a broad energy distribution that strongly impacts their transport properties and plasma chemistry. Fluid and hybrid models either assume a Maxwellian electron energy distribution, solve Boltzmann equation to determine electron energy distribution or use Monte Carlo techniques. Fluid plasma models have been coupled to models of external driving circuitry,[160] kinetic models that compute quantities not well captured by fluid models (e.g., ion and neutral energy and angular distribution at surfaces),[161] and surface physics models.[162] It is fair to state that plasma equipment modeling is a mature area and plasma reactor dynamics can be simulated with reasonable fidelity. Uncertainty with plasma chemistry[5] (atomic and molecular processes,

heavy particle reactions) is often the only hurdle that hinders the use of plasma equipment models for an even broader set of applications. Plasma equipment models have been successfully applied to the modeling of capacitively coupled plasma etchers (single frequency, dual frequency,[163] magnetized[164]), ICP sources,[157] ECR,[165] and helicon[166] plasmas. Fair to adequate mechanisms exist for many of the commonly used plasma etching gases.

Feature scale models have immensely grown in maturity in the last few years. Several techniques have been explored for feature scale modeling and they all remain equally important for problem solving. Broadly speaking, feature scale modeling has been done using Monte Carlo methods,[167] string based methods[168] and level set methods.[169] In Monte Carlo models, surface and material underneath is represented using macro-particles. Plasma species, whose characteristics are either assumed or determined using plasma equipment simulations, are then bombarded on the material stack. A surface reaction mechanism is used to determine how the structure evolves in time. Monte Carlo simulators allow representation of detailed surface processes and can easily account for sub-surface processes. However, to overcome the statistical noise in the simulations, large number of particles often have to be used slowing down simulations considerably. In string based methods, the surface of the structure is represented using a set of inter-connected strings in 2D models (or patches in 3D simulations). Using fluxes of plasma species, impingent flux on the structure surface is determined. Fluxes on material surface are used in conjunction with a surface mechanism to determine how the strings or patches evolve in time. String based techniques are computationally fast and it is relatively straight-forward to implement most surface processes. Representation of sub-surface material and simultaneous etching and deposition is nonetheless a non-trivial task in string based models. Level set methods have been used for both etching and deposition modeling. The material is represented by a function, one of whose equipotential planes coincides with the structure surface. Simulation methodology is similar to string based models although the surface is evolved by solving a differential equation governing the function. Level set method is slower than string based technique but simulations are numerically more stable. Representation of sub-surface materials and simultaneous etching and deposition are however challenges level set methods share with string based models. Feature scale models have been applied to the modeling of a wide variety of plasma etching processes including polysilicon[38] and photoresist[28] etching, and SiO_2[170] and low-κ[171] dielectric etching.

Molecular dynamics models have in recent years started playing a major role in unearthing the fundamentals of plasma etching. In the MD models, quantum mechanical interactions between atoms (both material and plasma based) are represented using pseudo-potentials that are either determined experimentally or using ab initio quantum mechanics models. These pseudo-potentials are used in classical mechanics models to simulation the dynamics on the material surface in contact with the plasma. Molecular dynamics models have been used to understand the formation of reactive layers on a variety of films, and the role that different ions and radicals play in plasma etching or surface passivation. These models have examined Cl_2[172] and fluorocarbon[173] etching of Si, fluorine etch of SiO_2[174] and fluorocarbon etching of SiO_2.[175]

References

1. International Technology Roadmap for Semiconductors, Semiconductor Industry Association, 2003 Edition.
2. Lieberman, M. A., and A. J. Lichtenberg. *Principles of Plasma Discharges and Materials Processing*. New York: Wiley, 1994.
3. Rauf, S. *J. Appl. Phys.* 87 (2000): 7647.
4. Wang, S.-B., and A. E. Wendt. *J. Vac. Sci. Technol. A* 19 (2001): 2425.
5. Christophorou, L. G., and J. K. Olthoff. *Fundamental Electron Interactions with Plasma Processing Gases*. London: Kluwer, 2004.

6. Nastasi, M., J. W. Mayer, and J. K. Hirvonen. *Ion Solid Interactions: Fundamentals and Applications*. Cambridge, MA: Cambridge University Press, 1996.
7. Eisele, K. M. *J. Electrochem. Soc.* 128 (1981): 123.
8. Coburn, J. W., and H. F. Winters. *J. Appl. Phys.* 50 (1979): 3189.
9. Standaert, T. E. F. M., M. Schaepkens, N. R. Rueger, P. G. M. Sebel, G. S. Oehrlein, and J. M. Cook. *J. Vac. Sci. Technol. A* 16 (1998): 239.
10. Atkins, P. *Physical Chemistry*, 5th ed., 877. New York: Freeman, 1994.
11. Bestwick, T. D., G. S. Oehrlein, and D. Angell. *Appl. Phys. Lett.* 57 (1990): 431.
12. Sekine, M. *Appl. Surf. Sci.* 192 (2002): 270.
13. Kitajima, T., Y. Takeo, and T. Makabe. *J. Vac. Sci. Technol. A* 17 (1999): 2510.
14. Keller, J. H., J. C. Forster, and M. S. Barnes. *J. Vac. Sci. Technol. A* 5 (1993): 2487.
15. Asmussen, J. Jr., T. A. Grotjohn, M. Pengun, and M. A. Perrin. *IEEE Trans. Plasma Sci.* 25 (1997): 1196.
16. Chen, F. F., and R. W. Boswell. *IEEE Trans. Plasma Sci.* 25 (1997): 1245.
17. Samukawa, S., K. Sakamoto, and K. Ichiki. *Jpn. J. Appl. Phys.* 40 (2001): L779.
18. Hwang, G. S., and K. P. Giapis. *J. Vac. Sci. Technol. B* 15 (1997): 70.
19. Tonnis, E. J., D. B. Graves, V. H. Vartanian, L. Beu, T. Lii, and R. Jewett. *J. Vac. Sci. Technol. A* 18 (2000): 393.
20. Armacost, M., P. D. Hoh, R. Wise, W. Yan, J. J. Brown, J. H. Keller, G. A. Kaplita, et al. *IBM J. Res. Dev.* 43 (1999): 39.
21. Chatterjee, A., I. Ali, K. Joyner, D. Mercer, J. Kuehne, M. Mason, E. Esquivel, et al. *J. Vac. Sci. Technol. B* 15 (1997): 1936.
22. Yeon, C.-K., and H.-J. You. *J. Vac. Sci. Technol. A* 16 (1998): 1502.
23. Ullal, S., H. Singh, V. Vahedi, and E. Aydil. *J. Vac. Sci. Technol. A* 20 (2002): 499.
24. Nagase, M., N. Ikezawa, K. Tokashiki, M. Takimoto, and K. Kasama. *Dry Process Symposium*, Tokyo, Japan, Paper I-3, 2001.
25. Bell, F. H., and O. Joubert. *J. Vac. Sci. Technol. B* 15 (1997): 88.
26. Greer, F., D. Fraser, J. W. Coburn, and D. B. Graves. *J. Appl. Phys.* 94 (2003): 7453.
27. Ling, L., X. Hua, X. Li, G. S. Oehrlein, E. A. Husdon, P. Lazzeri, and M. Anderle. *J. Vac. Sci. Technol. B* 22 (2004): 2594.
28. Rauf, S. *J. Vac. Sci. Technol. B* 22 (2004): 202.
29. DeBord, J. R. D., V. Jayaraman, M. Hewson, W. Lee, S. Nair, H. Shimada, V. L. Linh, J. Robbins, and A. Sivasothy. *2004 International Conference on Microelectronics and Interfaces Proceedings*, 229, 1998.
30. Kim, Y., J. Lee, H. Cho, and J. Moon. *2000 International Conference on Micro-process and Nanotechnology Proceedings*, 106, 2000.
31. Sato, Y., E. Shiobara, Y. Onishi, S. Yoshikawa, Y. Nakano, and S. Hayase. "Y Hamada." *J. Vac. Sci. Technol. B* 20 (2002): 909.
32. Lee, G. Y., Z. G. Lu, D. M. Dobuzinsky, X. J. Ning, and G. Costrini. *International Interconnect Technology Conference Proceedings*, 87, 1998.
33. Rauf, S., P. J. Stout, and J. Cobb. *J. Vac. Sci. Technol. B* 19 (2001): 172.
34. Goldfarb, D. L., A. Mahorowala, G. M. Gallatin, K. E. Petrillo, S. Ragson, H. H. Sawin, S. D. Allen, M. C. Lawson, and R. W. Kwong. *J. Vac. Sci. Technol. B* 22 (2004): 647.
35. Chan, V. W. C., C. H. Hai, and P. C. H. Chan. *J. Vac. Sci. Technol. B* 19 (2001): 743.
36. Xu, S., T. Lill, and D. Podlesnik. *J. Vac. Sci. Technol. A* 19 (2001): 2893.
37. Foucher, J., G. Cunge, L. Vallier, and O. Joubert. *J. Vac. Sci. Technol. B* 20 (2002): 2024.
38. Tuda, M., K. Shintani, and H. Ootera. *J. Vac. Sci. Technol. A* 19 (2001): 711.
39. Vallier, L., J. Foucher, X. Detter, E. Pargon, O. Joubert, and G. Cunge. *J. Vac. Sci. Technol. B* 21 (2003): 904.
40. Xu, S., Z. Sun, A. Chen, X. Quian, and D. Podlesnik. *J. Vac. Sci. Technol. A* 19 (2001): 871.
41. Hung, C.-C., H. C. Lin, M.-F. Wang, T.-Y. Huang, and H.-C. Shih. *Microelectron. Eng.* 63 (2002): 405.
42. Tuda, M., K. Shintani, and J. Tanimura. *Appl. Phys. Lett.* 79 (2001): 2535.

43. Vitale, S. A., and B. A. Smith. *J. Vac. Sci. Technol. B* 21 (2003): 2205.
44. Bell, F. H., O. Joubert, and L. Vallier. *J. Vac. Sci. Technol. B* 14 (1996): 96.
45. Bell, F. H., and O. Joubert. *J. Vac. Sci. Technol. B* 14 (1996): 2493.
46. Bell, F. H., and O. Joubert. *J. Vac. Sci. Technol. B* 15 (1997): 88.
47. Xu, S., Z. Sun, X. Quian, J. Holland, and D. Podlesnik. *J. Vac. Sci. Technol. A* 19 (2001): 166.
48. Kraft, R., I. Gupta, and T. Kinoshita. *J. Vac. Sci. Technol. B* 16 (1998): 496.
49. Desvoivres, L., L. K. Vallier, and O. Joubert. *J. Vac. Sci. Technol. B* 18 (2000): 156.
50. Panagopoulos, T., N. Gani, M. Shen, Y. Du, and J. Holland. In *Proceedings of the 2004 International Conference on Microelectronics and Interfaces*, 170, 2004.
51. Shan, H., B. Y. Pu, K.-H. Ke, M. Welch, and C. Deshpandey. *J. Vac. Sci. Technol. B* 14 (1996): 521.
52. Detter, X., R. Palla, I. Tomas-Boutherin, E. Pargon, G. Cunge, O. Joubert, and L. Vallier. *J. Vac. Sci. Technol. B* 21 (2003): 2174.
53. Gambino, J. P., and E. G. Colgan. *Mater. Chem. Phys.* 52 (1998): 99.
54. Sekine, M., Y. Kakuhara, and T. Kikkawa. *Electron. Commun. Jpn, Part 2* 79 (1996): 93.
55. Schreyer, T. A., A. J. Bariya, J. P. McVittie, and K. C. Saraswat. *J. Vac. Sci. Technol. A* 6 (1988): 1402.
56. Ikeguchi, K., J. Zhang, H. Lee, and C-L. Yang. In *Proceedings of the 2004 International Conference on Microelectronics and Interfaces*, 98, 2004.
57. Kim, J. S., W. T. Tang, W. S. Lee, B. Y. Yoo, Y. C. Shin, T. H. Kim, K. Y. Lee, Y. J. Park, and J. W. Park. *J. Vac. Sci. Technol. B* 17 (1999): 2559.
58. Ohiwa, T., A. Kojima, M. Sekine, I. Sakai, S. Yonemoto, and Y. Watanabe. *Jpn. J. Appl. Phys.* 37 (1998): 5060.
59. Matsui, M., T. Tatsumi, and M. Sekine. *Plasma Sources Sci.Technol.* 11 (2002): A202.
60. Zhang, Y., G. S. Oehrlein, and F. H. Bell. *J. Vac. Sci. Technol. A* 14 (1996): 2127.
61. Schaepkens, M., G. S. Oehrlein, C. Hedlund, L. B. Jonsson, and H.-O. Blom. *J. Vac. Sci. Technol. A* 16 (1998): 3281.
62. Givens, J., S. Geissler, J. Lee, O. Cain, J. Marks, P. Keswick, and C. Cunningham. *J. Vac. Sci. Technol. B* 12 (1994): 427.
63. Shon, J., T. Chien, H. Y. Kim, W. S. Lee, J. Kim, and D. Kiel. In *Proceedings of the 2001 Dry Process Symposium*, 225, Paper VI-7, 2001.
64. Kim, J., C. W. Chu, C. J. Kang, W. S. Han, and J. T. Moon. *J. Vac. Sci. Technol. B* 20 (2002): 2065.
65. Quiao, J., B. Jin, P. Phatak, J. Yu, and S. Geha. *J. Vac. Sci. Technol. B* 17 (1999): 2373.
66. Kim, J.-H., J.-S. Yu, C.-K. Ryu, S.-J. Oh, S.-B. Kim, J.-W. Kim, J.-M. Hwang, S.-Y. Lee, and I. Kouichiro. *J. Vac. Sci. Technol. A* 18 (2000): 1401.
67. Sun, Y.-C., and T.-Y. Huang. In *Proceedings of the Dry Process Symposium*, 219, Paper VI-6, 2001.
68. Ito, S., H. Namba, T. Hirata, K. Ando, S. Koyama, N. Ikezawa, T. Suzuki, T. Saitoh, and T. Horiuchi. *Microelectron. Reliab.* 42 (2002): 201.
69. Hook, T. B., D. Harmon, and C. Lin. *Microelectron. Reliab.* 41 (2001): 751.
70. Brozek, T., J. Huber, and J. Walls. *Microelectron. Reliab.* 40 (2000): 625.
71. Chang, K.-M., T.-H. Yeh, I.-C. Deng, and H.-C. Lin. *J. Appl. Phys.* 80 (1996): 3048.
72. Lin, B.-W.S.-S., C.-S. Tsai, and C.-C. Hsia. *Microelectron. Reliab.* 40 (2000): 2039.
73. Hashimoto, K., F. Shimpaku, A. Hasegawa, Y. Hikosaka, and M. Nakamura. *Thin Solid Films* 316 (1995): 1.
74. Lee, Y. J., S. W. Hwang, G. Y. Yeon, J. W. Lee, and J. Y. Lee. *Thin Solid Films* 341 (1999): 168.
75. Keil, D. L., B. A. Helmer, and S. Lassig. *J. Vac. Sci. Technol. B* 21 (2003): 1969.
76. Kropewnicki, T., K. Doan, B. Tang, and C. Björkman. *J. Vac. Sci. Technol. A* 19 (2001): 1384.
77. Mogab, C. J., A. C. Adams, and D. L. Flamm. *J. Appl. Phys.* 49 (1978): 3796.
78. Tatsumi, T., M. Matsui, M. Okigawa, and M. Sekine. *J. Vac. Sci. Technol. B* 18 (2000): 1897.
79. Matsui, M., T. Tastsumi, and M. Sekine. *J. Vac. Sci. Technol. A* 19 (2001): 1282.
80. Kurihara, K., Y. Yamaoka, K. Karahashi, and M. Sekine. *J. Vac. Sci. Technol. A* 22 (2004): 2311.
81. Stoffels, W. W., E. Stoffels, and K. Tachibana. *J. Vac. Sci. Technol. A* 16 (1998): 87.
82. Standaert, T. E. F. M., C. Hedlund, E. A. Joseph, G. S. Oehrlein, and T. J. Dalton. *J. Vac. Sci. Technol. A* 22 (2004): 53.
83. Doh, H.-H., J. H. Kim, S.-H. Lee, and K.-W. Whang. *J. Vac. Sci. Technol. A* 14 (1996): 2827.

84. Li, X., X. Hua, M. Fukusawa, and G. S. Oehrlein. *J. Vac. Sci. Technol. A* 21 (2003): 284.
85. Teii, K., M. Hori, M. Ito, and T. Goto. *J. Vac. Sci. Technol. A* 18 (2000): 1.
86. Hori, M., and T. Goto. *Appl. Surf. Sci.* 192 (2002): 135.
87. Li, X., X. Hua, M. Fukusawa, and G. S. Oehrlein. *J. Vac. Sci. Technol. A* 21 (2003): 284.
88. Tsukada, T., H. Nogami, Y. Nakagawa, E. Wani, K. Mashimo, H. Sato, and S. Samukawa. *Thin Solid Films* 341 (1999): 84.
89. Hua, X., C. Stolz, G. S. Oehrlein, P. Lazzeri, N. Coghe, M. Anderle, C. K. Inoki, T. S. Kuan, and P. Jiang. *J. Vac. Sci. Technol. A* 23 (2005): 151.
90. Posseme, N., O. Joubert, L. Vallier, and N. Rochat. *J. Vac. Sci. Technol. B* 22 (2004): 2772.
91. Sankaran, A., and M. J. Kushner. *Appl. Phys. Lett.* 82 (2003): 1824; Sankaran, A., and M. J. Kushner. *J. Vac. Sci. Technol. A* 22 (2004): 1260.
92. Min, J.-H., S.-W. Hwang, G.-R. Lee, and S.-H. Moon. *J. Vac. Sci. Technol. B* 21 (2000): 1210.
93. Kojima, A., T. Sakai, and T. Ohiwa. *J. Vac. Sci. Technol. B* 22 (2004): 2611.
94. Claeys, C. In *Proceedings of the 17th IEEE Conference VLSI Design* (*VLSID '04*), 275, 2004.
95. Onai, T., S. Kimura, K. Suko, H. Miyazaki, and F. Yano. *Hitachi Rev.* 52 (2003): 117.
96. Ito, T. *Fujitsu Sci. Tech. J.* 39 (2003): 3.
97. Hiramoto, T. *2004 IEEE Conference on Integrated Circuit Design and Technology*, 59, 2004.
98. Samavedam, S., L. La, J. Smith, S. Dakshina-Murphy, E. Luckoski, J. Schaeffer, M. Zalava, et al. *2002 Symposium on VLSI Technology*, 24, 2002.
99. Nguyen, B.-Y., A. Thean, T. White, A. Vandooren, M. Sadaka, L. Mathew, A. Barr, et al. *2004 IEEE Conference on Integrated Circuit Design and Technology*, 237, 2004.
100. Chang, L., C. Yang-kyu, D. Ha, P. Ranade, X. Shiying, J. Bokor, H. Chenming, and T. J. King. *Proc. IEEE* 91 (2003): 1860.
101. Chuang, C.-T., K. Bernstein, R. V. Joshi, R. Puri, K. Kim, E. J. Nowak, T. Ludwig, and I. Aller. *IEEE Circuits Dev.* 20, no. 1 (2004): 6.
102. Englehart, M., G. Schindler, W. Steinhögl, and G. Steinlesberger. *Microelectron. Eng.* 64 (2002): 3.
103. Rottstegge, J., W. Herbst, S. Hien, G. Futterer, C. Eshbaumer, C. Hohle, J. Schwider, and M. Sebald. *Proc. SPIE* 4690 (2004): 233.
104. Lin, B. J. *Proc. SPIE* 5377 (2004): 46.
105. Switkes, M., R. R. Kunz, M. Rothschild, R. F. Sinta, M. Yeung, and S.-Y. Baek. *J. Vac. Sci. Technol. B* 21 (2003): 2794.
106. Zandbergen, P., D. Van Steenwinckel, J. H. Lammers, H. Kwinten, and C. Juffermans. *IEEE Trans. Semi. Manufact.* 18 (2005): 37.
107. Tserepi, A., G. Cordoyiannis, G. P. Patsis, V. Constantoudis, E. Gogolides, E. S. Valamontes, D. Eon, et al. *J. Vac. Sci. Technol. B* 21 (2003): 174.
108. Yoon, J.-Y., M. Hata, J.-H. Hah, H.-W. Kim, S.-G. Woo, H.-K. Cho, and W.-S. Han. *Proc. SPIE* 5376 (2004): 196.
109. Foucher, J., G. Cunge, L. Vallier, and O. Joubert. *J. Vac. Sci. Technol. B* 20 (2002): 2024.
110. Constantoudis, V., G. Patsis, and E. Gogolides. *Proc. SPIE* 5038 (2003): 901.
111. Satou, I. *Jpn. J. Appl. Phys.* 38 (1999): 7008.
112. Yoshizama, M., S. Moriya, H. Nakano, Y. Shirai, T. Morita, T. Kitagawa, and Y. Miyamoto. *Jpn. J. Appl. Phys.* 43 (2004): 3739.
113. Vandooren, A., A. Barr, L. Mathew, T. R. White, S. Egley, D. Pham, M. Zavala, et al. *IEEE Elect. Dev. Lett.* 24 (2003): 342.
114. Shimada, H., and K. Maruyama. *Jpn J. Appl. Phys.* 43 (2004): 1768.
115. Sha, L., and J. P. Chang. *J. Vac. Sci. Technol. A* 22 (2004): 88.
116. Monfray, S., A. Souifi, F. Boeuf, C. Ortolland, A. Poncet, L. Militaru, D. Chanemougame, and T. Skotnicki. *IEEE Trans. Nanotechnol.* 2 (2003): 295.
117. Nowak, E. J., I. Aller, T. Ludwig, K. Kim, R. V. Joshi, C.-T. Chuang, K. Bernstein, and R. Puri. *IEEE Circuits Dev.* 20 (2004): 20.
118. Schaeffer, J., C. Capasso, L. Foneca, S. Samavedam, D. Gilmer, Y. Liang, S. Kalpat, et al. *2004 IEEE Int. Elect. Dev. Mtg.* 287, (2004).

119. Wong, H.-S., B. Doris, E. Gusev, M. Icong, E. Jones, J. Kedzieski, Z. Ren, K. Rim, and H. Shang. *2003 IEEE Symposium on VLSI Technology*, 13, 2003.
120. Matthews, L., D. Yang, A. Thean, A. Vandooren, C. Parker, T. Stephens, R. Mora, et al. *IEEE Symposium on VLSI Technology*, 97, 2005.
121. Gossmann, H.-J. L., A. Agarwal, T. Parrill, L. M. Rubin, and J. M. Poate. *IEEE Trans. Nanotechnol.* 2 (2003): 285.
122. Borel, S., C. Arvet, J. Bildie, V. Caubet, and D. Louis. *Jpn. J. Appl. Phys.* 43 (2004): 3964.
123. Sparks, T., S. Rauf, L. Vallier, G. Cunge, and T. Chevolleau. *International Symposium of AVS*, 2004.
124. Liang, M.-C., H.-Y. Tsai, C.-C. Chung, C.-C. Hsueh, H. Chung, and C.-Y. Lu. *IEEE Elect. Dev. Lett.* 24 (2003): 562.
125. Case, C. *Future Fab Int.* 17, (2004) chap. 6.
126. Matsumoto, S., A. Ishii, K. Hashimoto, Y. Nishioka, M. Sekiguchi, S. Isono, T. Satake, et al. *IEEE International Interconnect Technology Conference* (*IITC*), 262, 2003.
127. Moore, D., R. Carter, H. Cui, P. Burke, P. McGrath, S. Q. Gu, D. Gidley, and H. Peng. *J. Vac. Sci. Technol. B* 23 (2005): 332.
128. Maex, K., M. R. Baklanov, D. Shamiryan, F. lacopi, S. H. Brongersma, and Z. S. Yanovitskaya. *J. Appl. Phys.* 93 (2003): 8793.
129. Steinlesberger, G., M. Engelhart, G. Schindler, W. Steinhogl, A. von Glasow, and K. Mosig. Impact of Annealing on the Resistivity of Ultrafine Cu Damascene Interconnect. *Mat. Tech. and Rel. for Adv. Interconnect. Symp. Proceedings*, 766, 2003.
130. Connelly, D., C. Faulkner, D. E. Grupp, and J. S. Harris. *IEEE Trans. Nanotechnol.* 3 (2004): 98.
131. Goronkin, H., P. Von Allmen, R. Tsui, and T. Zhu. *Nanostruct. Sci. Technol.* 67, (1999), chap. 5.
132. De Blauwe, J. *IEEE Trans. Nanotechnol.* 1 (2002): 72.
133. Ma, Y., G. Tavid, and F. Cerrina. *J. Vac. Sci. Technol. B* 22 (2004): 3124.
134. Chan, V., C. Hai, and P. Chan. *J. Vac. Sci. Technol. B* 19 (2001): 743.
135. Finch, R., J. Feldman, F. Zumsteg, M. Crawford, A. Feiring, V. Petrov, F. Schadt III., and R. Wheland. *Semicond. Fabtech.* 14 (2001): 167.
136. Shenoy, R., and K. Saraswat. *IEEE Trans. Nanotechnol.* 2 (2003): 265.
137. Yang, F.-L., D.-H. Lee, H.-Y. Chen, C.-Y. Chang, S.-D. Liu, C.-C. Huang, et al. *IEEE Symposium on VLSI Technology*, 196, 2004.
138. Hu, S.-F., Y.-C. Wu, C.-L. Sung, C.-Y. Chang, and T.-Y. Huang. *IEEE Trans. Nanotechnol.* 3 (2004): 93.
139. Van den hove, L., A. Goethals, K. Ronse, M. Van Bavel, and G. Vandenberghe. *IEEE Int. Elect. Dev. Mtg.* 3, (2002).
140. Molas, G., B. De Salvo, G. Ghibaudo, D. Mariolle, A. Toffoli, N. Buffet, R. Puglisi, S. Lombardo, and S. Deleonibus. *IEEE Trans. Nanotechnol.* 3 (2004): 42.
141. Sano, N., A. Hiroki, and K. Matsuzawa. *IEEE Trans. Nanotechnol.* 1 (2002): 63.
142. Zhuang, L., L. Guo, and S. Y. Chou. *Appl. Phys. Lett.* 72 (1998): 1205.
143. Engelhardt, M., G. Schindle, W. Steinhogl, and G. Steinlesberger. *Microelectron. Eng.* 64 (2002): 3.
144. Milanovic, V., and L. Doherty. In *Proceedings of ASME International Mechanical Engineering Congress and Exposition* (*IMECE*), 33392, 2002.
145. Noda, S., H. Nishimori, T. Ida, T. Arikado, K. Ichiki, T. Ozaki, and S. Samukawa. *J. Vac. Sci. Technol. A* 22 (2004): 1507.
146. Kim, S. J., H. J. Lee, G. Y. Yeom, and J. K. Lee. *Jpn. J. Appl. Phys.* 43 (2004): 7261.
147. Sommervell, M., D. Fryer, B. Osburn, K. Patterson, J. Byers, and C. Wilson. *J. Vac. Sci. Technol. B* 18 (2000): 2551.
148. Brown, K. *Future Fab Int.* 17, (2004).
149. Litt, L., B. Roman, and J. Cobb. *Future Fab Int.* 17, (2004).
150. Naeemi, A., and J. Meindl. *IEEE Elect. Dev. Lett.* 26 (2005): 84.
151. Kreup, F., A. Graham, M. Liebau, G. Duesberg, R. Seidel, and E. Unger. *IEEE Int. Elect. Dev. Mtg.* 683, (2004).
152. Yang, J. *IEEE Circuits Dev.* 1 (2004): 44.
153. Lieberman, M. A. *IEEE Trans. Plasma Sci.* 16 (1988): 638.

154. Birdsall, C. K. *IEEE Trans. Plasma Sci.* 19 (1991): 65.
155. White, R. D., K. F. Ness, and R. E. Robson. *Appl. Surf. Sci.* 192 (2002): 26.
156. Graves, D. B., and K. F. Jensen. *IEEE Trans. Plasma Sci.* 14 (1986): 78.
157. Ventzek, P. L. G., R. J. Hoekstra, and M. J. Kushner. *J. Vac. Sci. Technol. B* 12 (1994): 461.
158. http://www.plasmator.com
159. http://www.cfdrc.com
160. Rauf, S., and M. J. Kushner. *J. Appl. Phys.* 83 (1998): 5087.
161. Hoekstra, R. J., and M. J. Kushner. *J. Appl. Phys.* 79 (1996): 2275.
162. Zhang, D., and M. J. Kushner. *J. Appl. Phys.* 87 (2000): 1060.
163. Wakayama, G., and K. Nanbu. *IEEE Trans. Plasma Sci.* 31 (2003): 638.
164. Rauf, S. *IEEE Trans. Plasma Sci.* 31 (2003): 471.
165. Kinder, R., and M. J. Kushner. *J. Vac. Sci. Technol. A* 17 (1999): 2421.
166. Kinder, R. L., A. R. Ellingboe, and M. J. Kushner. *Plasma Sources Sci. Technol.* 12 (2003): 561.
167. Hoekstra, R. J., M. J. Grapperhaus, and M. J. Kushner. *J. Vac. Sci. Technol. A* 15 (1997): 1913.
168. Abdollahi-Alibeik, S., J. P. McVittie, K. C. Saraswat, V. Sukharev, and P. Schoenborn. *J. Vac. Sci. Technol. A* 17 (1999): 2485.
169. Hwang, H. H., T. R. Govindan, and M. Meyyappan. *J. Electrochem. Soc.* 146 (1999): 1889.
170. Zhang, D., S. Rauf, T. Sparks, and P. L. G. Ventzek. *J. Vac. Sci. Technol. B* 21 (2003): 828.
171. Sankaran, A., and M. J. Kushner. *J. Vac. Sci. Technol. A* 22 (2004): 1242.
172. Barone, M. E., and D. B. Graves. *J. Appl. Phys.* 77 (1995): 1263.
173. Abrams, C. F., and D. B. Graves. *J. Appl. Phys.* 86 (1999): 5938.
174. Ohta, H., and S. Hamaguchi. *J. Vac. Sci. Technol. A* 19 (2001): 2373.
175. Smirnov, V. V., A. V. Stengach, K. G. Gaynullin, V. A. Pavlovsky, S. Rauf, P. J. Stout, and P. L. G. Ventzek. *J. Appl. Phys.* 97 (2005): 093302.

22

Equipment Reliability

22.1 Introduction **22**-1
Basic Definitions • Reliability of a Repairable System • Reliability Metrics

22.2 Reliability Metrics Calculations **22**-4

22.3 Applications of Reliability Metrics **22**-5
Desired Values • Analytical/Theoretical Values • Observed Values

22.4 Confidence Limits Calculations **22**-6
Confidence Limit Calculations for Theoretical Values • Confidence Limit Calculations for Observed Values

22.5 Precise Use of the Reliability Metrics **22**-7

22.6 Maintainability Metrics **22**-8

22.7 High-Level Equipment Performance Metrics **22**-8
Availability (Uptime) and Utilization Metrics • Overall Equipment Efficiency • Cost of Ownership • Hierarchy of Equipment Performance Metrics

22.8 An Example of Reliability and High Level Performance Metrics Calculations **22**-11
Given Values • Metrics Calculations

22.9 Four Steps to Better Equipment Reliability **22**-12
Know Goals and Requirements • Design-In Reliability • Build-In Reliability • Manage Reliability Growth

22.10 Reliability Testing **22**-20
Types of Reliability Tests • Generic Steps for Reliability Tests • Reliability Tests throughout the Equipment Program Life Cycle Phases

22.11 Use of Equipment Reliability Discipline in Business Practices **22**-24

22.12 SEMI E10 ..... **22**-25
Benefits of Using SEMI E10 • Key Elements of SEMI E10

References **22**-27

Vallabh H. Dhudshia
SafeFab Solutions

22.1 Introduction

Reliability has been widely used to measure equipment performance in military and commercial industry since the early 1940s. Since then, the importance of the reliability has grown at a phenomenal rate. Now, reliability is a key equipment characteristic that has significant influence over equipment production efficiency and cost of owning and operating a piece of equipment. In addition, better reliability leads to a

competitive advantage. In the semiconductor manufacturing equipment industry, reliability plays even greater role enabling semiconductor manufacturers to compete globally.

Currently, some semiconductor manufacturers also track high-level metrics, such as overall equipment efficiency (OEE) or cost of ownership (CoO), for equipments in their factories. Reliability is a key element of such metrics.

This chapter provides basic working knowledge of equipment reliability discipline, how to define, how to calculate reliability and related metrics, and their inter dependencies. Also, included are their application to semiconductor manufacturing operations and ways to improve them.

22.1.1 Basic Definitions

22.1.1.1 Equipment

Equipment is defined as a combination of components, parts, assemblies, modules, accessories, and embedded software integrated to perform the intended functions. At least one component must fail to cause the equipment to fail to perform its intended functions. Most semiconductor manufacturing equipments are repairable systems. A repairable system is one, which after failing to perform at least one of its intended functions, can be restored to perform all of its intended functions by any method other than replacing the entire system. A repairable system can be restored by replacing, repairing, adjusting, or cleaning the appropriate component(s); and/or reloading embedded software. All the subject matter presented here refers to such repairable systems, either semiconductor manufacturing equipment or other repairable systems.

22.1.1.2 Reliability

Equipment has many characteristics of interest to us, such as weight, volume, footprint, throughput rate, price, etc. One of the important characteristics is reliability as defined below.

Reliability is a longevity measure of failure-free operation period of any equipment. It can be expressed in many different ways. One way is using the following formal definition.

Reliability is the probability of performing intended functions continuously without a failure (stoppage or interruption) for a specified time under the stated operational conditions.

Mathematically, it is written as:

$$R(t) = \Pr[T > t] \tag{22.1}$$

where t, specific time of interest; T, random variable; $R(t)$, reliability at time t; Pr[], probability of.

Example:

$$R(1000\text{ h}) = \Pr[T > 1000\text{ h}] = 0.95 \tag{22.2}$$

In this example, 95% of the equipment units should continue operations past 1000 h.

Four key points of the above formal definition require further explanation.

22.1.1.2.1 Failure

A failure is defined as any unscheduled event that changes the equipment to a condition, where it cannot perform one of its intended functions. Any part failure, out of adjustment or contaminated parts, software or process recipe problem, facility or utility supply malfunction, or human error could cause the failure.

22.1.1.2.2 Intended Functions

Every piece of equipment has its intended functions, whether they are formally documented or not. However, a given reliability level applies to a given set of functions that the equipment was designed to accomplish. If the equipment is used for functions other than its intended design, the same reliability

level may not apply to these new functions. It is the manufacturer's responsibility to see that users understand equipment's intended functions and vice versa.

22.1.1.2.3 Specified Time

The reliability level changes as the equipment ages. It is necessary to include equipment age in establishing a reliability level. Without inclusion of such a time element, any reliability level is ambiguous. Such situations can mislead a user about the specific reliability level.

22.1.1.2.4 Stated Operational Conditions

Factors such as operating environment, operating stress level, operating speed, operator skill level, and maintenance procedures and policies can affect the reliability of any equipment. Therefore, a given reliability level applies to a specified set of the operational conditions. If the value of any factor varies from specified operational conditions, the reliability level may differ.

Example: the reliability of a blower in a card cage, operating in its ambient environment at 60% of its rated power, will be 0.85 at 2 years after installation.

22.1.2 Reliability of a Repairable System

In a repairable system, the distribution of failure intervals between successive failures is of prime interest. If we assume that:

1. Each component failure is an independent renewal process, i.e., when a component fails, it is replaced by a new component and this does not affect any other components
2. The system is a series system with many independent components.

Then, under very generic conditions, the system-level failure is a superimposed renewal process [1]. The time between two successive failures approximately follows an exponential distribution as shown below.

$$f(t) = e^{-(t/\theta)} = \lambda e^{-\lambda t} \tag{22.3}$$

where $f(t)$, probability density function (PDF) at time t; θ, mean life; λ, failure rate = 1/mean life.

Note that the failure rate is constant for exponential PDF. The reliability function for the exponential PDF is given by:

$$R(t) = e^{-(t/\theta)} = e^{-\lambda t} \tag{22.4}$$

The exponential distribution is one of the most popular, simple, and widely used PDF in reliability discipline. This approximation makes reliability analysis of a repairable system very easy. We need to know only one parameter (either mean time between failures (MTBF) or failure rate λ) of the distribution to be able to perform system-level reliability analyses.

22.1.3 Reliability Metrics

The *reliability metrics* are various terms used to quantify the numerical value of reliability levels. In semiconductor manufacturing industry, we use neither reliability $R(t)$ nor failure rate (λ) to measure level of equipment reliability. Instead, we use metrics based on mean life. These metrics consist of at least four words, as shown in Figure 22.1. Two of them, *MEAN* and *BETWEEN*, are mandatory. Other words relate to the measures of life and events.

Using the algorithm given in Figure 22.1, we can define metrics appropriate for any situation. Take the word *MEAN*, select a word for measure of life; take the word *BETWEEN*, and select the desired event.

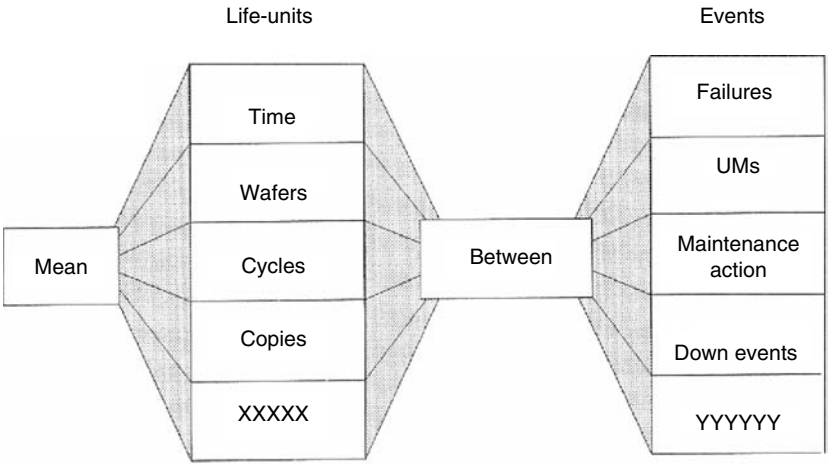


FIGURE 22.1 Reliability metrics algorithm.

Examples

1. Mean time between failures
2. Mean cycles between unscheduled maintenance
3. Mean wafers between failures.

These metrics are widely used to track reliability of semiconductor manufacturing equipment and are recommended by the Semiconductor Equipment Manufacturing International (SEMI), an association of semiconductor manufacturing equipment suppliers, in SEMI Specification E10-0304^{E} for definition and measurement of equipment reliability, availability, and maintainability (RAM) [2].

22.2 Reliability Metrics Calculations

To calculate the numerical value of equipment reliability metrics, we need to know two basic elements of the reliability discipline, (a) number of events of interest that stops equipment from performing its intended function(s) in a given life period, and (b) amount of life-units the equipment was in operational condition during the same period. Use the following equation to calculate the desired metric.

$$\text{Reliability metric} = \frac{\text{Operational life period}}{\text{Number of events during the same operational life period}} \tag{22.5}$$

As shown in Figure 22.1, semiconductor manufacturing events are categorized as failures, down events, scheduled maintenance or unscheduled maintenance. Operational life period may be expressed in calendar time, productive time, number of cycles, or number of wafer processed. Based on these variations, some popular reliability metrics for semiconductor manufacturing equipment are given by:

$$\text{Mean(productive) time between failures } (\text{MTBF}_p) = \frac{\text{Productive time}}{\text{Number of failures that occur during productive time}} \tag{22.6}$$

$$\text{Mean cycles between failures (MCBF)} = \frac{\text{Total equipment cycles}}{\text{Number of failures}} \tag{22.7}$$

The above formulae also apply to any measure of life by replacing "cycles" with the desired measures. For example,

$$\text{Mean wafers between failures (MWBF)} = \frac{\text{Number of wafers processed}}{\text{Number of failures}} \tag{22.8}$$

Section 22.8 contains an example of calculations for the above metrics.

22.3 Applications of Reliability Metrics

Applications of reliability metrics are terms that originate from the reliability related activities and they are used with appropriate metrics. For example, *reliability goal* originates from a goal-setting activity. *Goal* is an application of any reliability metric, e.g., goal MTBF.

The applications of these metrics can be divided into the following three categories, depending upon the origin of the activities they represent:

1. Desired values
2. Analytical/theoretical values
3. Observed values.

22.3.1 Desired Values

In this category, the value of metric levels originates from the activities that deal with desires of equipment manufacturers and users.

For example:

1. The MTBF goals are what a manufacturer (or a user) wants his equipment to perform.

Therefore, goal applications belong to the *Desired Value* category.

When system-level goals or requirements are broken into department level goals or requirements, based on some logical justification, they generate the applications known as allocation, budgeting, or apportionment, which also belong to the *Desired Value* category.

System-level reliability goals, allocated to subsystem and component, and corresponding operating environments are part of the respective design specifications for reliability. Therefore, design specification is also an application of the reliability metrics and falls in this category.

22.3.2 Analytical/Theoretical Values

In this category of applications, the value of metric levels originates from appropriate theoretical activities, such as modeling, part count calculation, and stress analysis.

Following are two typical examples of the *Analytical/Theoretical Values* applications:

1. *Inherent reliability*. The values are derived from design assessment, assuming benign environments and no error in design, manufacturing, and operation. The inherent reliability is the best achievable level.
2. *Expected reliability*. The values are based on theoretical calculations using theoretical or observed reliability level of the parts used in the equipment.

22.3.3 Observed Values

This category represents situations in which the metrics level is established based on actual in-house tests, field tests, or field operations of the equipment.

The following are three typical examples:

1. *Observed values*. The values are derived from actual in-house tests, field tests, or field operations. These values are not altered or adjusted.
2. *Assessed/Adjusted values*. When observed values are adjusted to account for non-relevant failures (such as facility problems or out-of-spec consumable), they become assessed/adjusted values.
3. *Confidence limit values*. These values are the observed values adjusted to account for number of failures observed as described in Section 22.4.

Figure 22.2 below shows the most commonly used terms for reliability metrics and their applications. The reliability metrics can be converted from one category to another. For example, if we know MTBF, we can convert it to failures per 1000 h. Applications of the metrics can neither be converted from one to another nor mixed. However, they can be compared (for example, goal value vs. observed values).

22.4 Confidence Limits Calculations

When we deal with reliability metrics, either analytical or observed, we always face a question, how much confidence do we have in the result. It varies depending upon the type of theoretical calculations made (to derive theoretical values) or number of failures observed and amount of productive time (or other measures of life) contained within the observed period. The confidence is expressed by calculating confidence limits of the calculated or observed values. Generally we are interested in lower confidence limit. Therefore, the calculation methods shown below show lower limit calculations only. Similar methodology is used to calculate upper confidence limits.

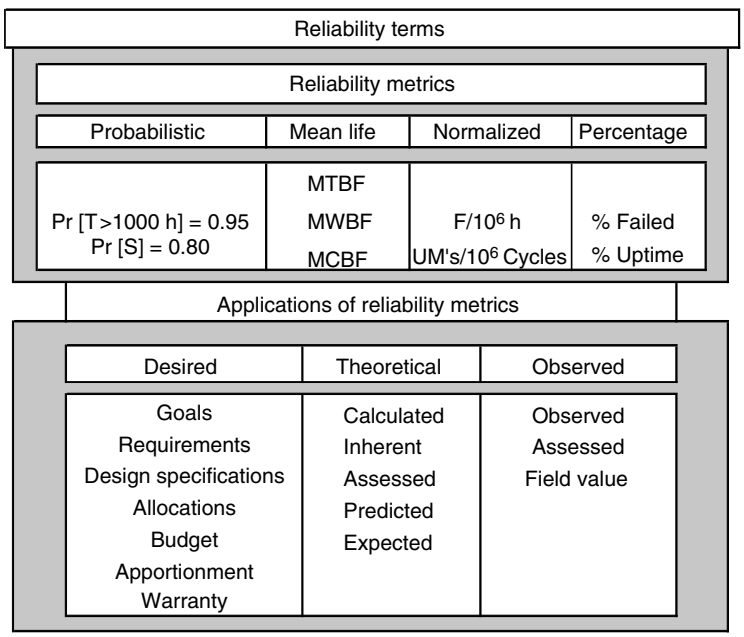


FIGURE 22.2 Summary of reliability metrics and their applications.

22.4.1 Confidence Limit Calculations for Theoretical Values

To calculate the confidence limit, we must have an underlying probability distribution for the calculated values. The underlying distribution depends upon the distribution of the MTBF or other values of the parts used in the calculations. Applying a law of large numbers, this distribution could be a normal distribution with a mean of the calculated values (μ), for the reliability metric under consideration, and their standard deviation (σ). In such cases, use the formulae given in Table 22.1 to calculate the lower confidence limit.

For example, if μ and σ of the calculated MTBF values are 500 and 100 h respective, then

$$80\% \text{ lower confidence limit for calculated MTBF} = 500 - 0.842 \times 100 = 415.8 \text{ h.} \tag{22.9}$$

22.4.2 Confidence Limit Calculations for Observed Values

Formulae given in Section 22.2 for reliability metrics provide single value estimates for the observed performance. The lower confidence limit for any observed reliability metric ($MTBF_p$, MWBF, MCBF, etc.,) depends upon number of failures observed and amount of productive time (or other measures of life) contained within the observed period. It is calculated using the following formula.

$$P\% \text{ lower confidence limit for the observed metrics} = (\text{observed value}) \times (\text{appropriate factor for the confidence level}) \tag{22.10}$$

For the semiconductor equipment industry, any value between 80 and 95% confidence levels is an accepted norm for reliability work. See Table 22.2 for the multiplier factor values at 80, 90, and 95% confidence levels [2].

Section 22.8 contains an example of lower confidence limit calculations.

22.5 Precise Use of the Reliability Metrics

We need the following items to fully and precisely define the reliability level of a real-life situation.

1. Appropriate application (e.g., goal)
2. Appropriate reliability metric (e.g., MTBF)
3. Appropriate numerical value (e.g., 5000)
4. Appropriate unit for the metric (e.g., hours)
5. Appropriate set of intended functions (e.g., metal etch process)
6. Age of the equipment when the metric and value apply (e.g., two months after installation)
7. Appropriate set of operational conditions (e.g., clean room environment)
8. Appropriate confidence level (e.g., 80%) for confidence limit values.

For examples:

1. Goal MTBF is 5000 h, 2 months after installation, when used as metal etcher in clean room environment.
2. Part count calculations for poly etcher model E3000 shows inherent failure rate of 2.00 failures per 1000 h (MTBF = 500 h) under the assumptions listed (needs to have a list of all assumptions used

TABLE 22.1 Formulae for Calculating Lower Confidence Limit of Theoretical Values

Confidence Limit	Formula Used
70% Lower confidence limit	$\mu - 0.525\sigma$
80% Lower confidence limit	$\mu - 0.842\sigma$
90% Lower confidence limit	$\mu - 1.282\sigma$
95% Lower confidence limit	$\mu - 1.645\sigma$

TABLE 22.2 Multiplier Factors for the Lower Confidence Limit Calculations

Number of Failures	80% Confidence	90% Confidence	95% Confidence
1	0.621	0.434	0.334
2	0.668	0.514	0.422
3	0.701	0.564	0.477
4	0.725	0.599	0.516
5	0.744	0.626	0.546
10	0.799	0.704	0.637
15	0.828	0.745	0.685
20	0.846	0.772	0.717
30	0.870	0.806	0.759

in the calculations, such as operating conditions, part functions, etc.). Assuming $\sigma = 100$, 80% lower confidence limit for the calculated MTBF = 415.8 h.

3. Observed MCBF = 20,000 cycles. When the facility related failures are discounted, the adjusted MCBF = 30,000 cycles. Ninety percent lower confidence limit MCBF = 21,030 cycles (based on three equipment failures). Equipment is 6-month old and used as an oxide etcher.

22.6 Maintainability Metrics

There is no direct relationship between reliability and maintainability. Reliability deals with the operational life longevity and failures of equipment, while maintainability deals with restoring the equipment operation and the time it takes to restore it. However, both disciplines are complementary; both support high-level equipment performance metrics described in Section 22.7.

Formally, *maintainability* is the probability that the equipment will be restored to a specific operational condition (able to perform its all intended functions) within a specified period of time, when the maintenance is performed by personnel having specified skill levels and using prescribed procedures, resources, and tools. Maintenance can be either unscheduled or scheduled. One of the most popular measures of maintainability is mean time to repair mean time to repair (MTTR), given by:

$$\text{Mean time to repair (MTTR)} = \frac{\text{Total repair time}}{\text{Number of repair events}} \tag{22.11}$$

Repair time includes diagnosis, corrective actions, and verification tests, but not maintenance delays.

22.7 High-Level Equipment Performance Metrics

Recently, some semiconductor manufacturing equipment users have started using high-level equipment performance metrics to make equipment purchase decisions and improve equipment performance. In addition, these high-level equipment performance metrics are becoming increasingly important to compete in the global market because they satisfy customer's reliability requirements in an optimum manner. Equipment reliability is the key element of these high-level metrics. The proper level of reliability is the one that yields the optimal value of the high-level metric being considered.

Three most widely used high-level equipment performance metrics in semiconductor manufacturing are:

1. Availability (Uptime) and utilization
2. Overall equipment efficiency
3. Cost of ownership.

22.7.1 Availability (Uptime) and Utilization Metrics

Availability is a joint measure of reliability and maintainability. It is defined as the probability that equipment will be in a condition to perform its intended functions when required. Percentage uptime is one of the most widely used metrics for availability in semiconductor manufacturing. Since equipment down time can be attributed to either the equipment, equipment supplier, or the equipment users, the uptime calculations vary accordingly. The following three kinds of uptime calculation are used in semiconductor manufacturing: (a) equipment dependent, (b) supplier dependent, and (c) operational.

Equipment dependent uptime includes effect of down time caused by scheduled and unscheduled maintenance inherent with the equipment (design) and it is given by:

$$\text{Equipment dependent uptime (\%)} = \frac{\text{Equipment uptime} \times 100}{(\text{Equipment uptime} + \text{DT}_\text{E})} \tag{22.12}$$

where DT_E = Equipment dependent down times = (unscheduled repair time + unscheduled and scheduled time to change consumables and chemicals + product test time + preventive maintenance time).

Equipment uptime includes productive, engineering, and standby times. It does not include non-scheduled time such as holidays, shutdowns, non-working shifts, etc. [2].

Supplier-dependent uptime includes effects of all equipment dependent down times (DT_E) and maintenance delays caused by the equipment supplier. It is given by:

$$\text{Supplier-dependent uptime (\%)} = \frac{\text{Equipment uptime} \times 100}{(\text{Equipment uptime} + \text{DT}_\text{E} + \text{supplier caused maintenance delays})} \tag{22.13}$$

Operational uptime includes effects of all down time caused by the scheduled and unscheduled maintenance inherent with the equipment, maintenance delays caused by the equipment supplier and user, and any other down time caused by the equipment user (such as facility related down time). It is given by:

$$\text{Operational uptime (\%)} = \frac{\text{Equipment uptime} \times 100}{(\text{Equipment uptime} + \text{Equipment down time})} \tag{22.14}$$

The above uptime formulae do not include non-scheduled time, such as vacation, holidays, shutdowns, etc. Total utilization includes all time when the equipment is utilized productively. It measures the overall asset efficiency and is given by:

$$\text{Total utilization (\%)} = \frac{\text{Productive time} \times 100}{\text{Total time}} \tag{22.15}$$

22.7.2 Overall Equipment Efficiency

Overall equipment efficiency is the most recent high-level equipment performance metric. It was developed as an equipment effectiveness metric in Japan to measure the effectiveness of a manufacturing technique called total productive maintenance (TPM). Originally, it was called overall equipment effectiveness. Semiconductor Equipment Manufacturing International Metric Committee changed it to overall equipment effectiveness [3]. Semiconductor Equipment Manufacturing International and the American Institute of Total Productive Maintenance (AITPM) are currently the major sponsor of the OEE metric in the U.S.A.

Overall equipment efficiency is an all-inclusive metric of equipment productivity, i.e., it is based on reliability (MTBF), maintainability (MTTR), utilization (availability), throughput, and yield. All the above factors are grouped into the following three submetrics of equipment efficiency.

1. Availability
2. Performance efficiency
3. Rate of quality.

The three submetrics and OEE are mathematically related as follows:

$$\text{OEE (\%)} = \text{Availability} \times \text{Performance Efficiency} \times \text{Rate of Quality} \times 100 \tag{22.16}$$

Now let us look at each OEE sub-metric in more detail.

22.7.2.1 Availability

We have already defined availability in Section 22.7.2. We can use any uptime metric in this equation depending upon which OEE we are calculating. For example, equipment-dependent OEE calculations use equipment-dependent uptime and so forth.

22.7.2.2 Performance Efficiency

The performance efficiency is based on losses incurred from idling, minor stops, and equipment speed losses. It is given by:

$$\text{Performance efficiency} = \frac{\text{Theoretical CT} \times \text{actual PPH}}{\text{Actual CT} \times \text{theoretical PPH}} \tag{22.17}$$

where CT, cycle time; PPH, throughput rate in parts (units) per hour.

22.7.2.3 Quality Rate

The quality rate is a measure of output quality and is given by:

$$\text{Quality rate} = \frac{\text{Total part produce} - \text{number of rejects}}{\text{Total parts produced}} \tag{22.18}$$

where rejects are defined as any produced part that does not meet the production criteria.

22.7.2.4 Simple OEE

There is a simple and quick way to calculate OEE without going into elaborate calculations of the above three sub-metrics.

$$\text{Simple OEE (\%)} = \frac{(\text{Number of good units produced in } t \text{ calendar hours}) \times 100}{(t \times \text{theoretical PPH})} \tag{22.19}$$

Note that this value gives only rough estimate for the OEE. It does not give any indication of improvement activities direction. There are many other ways to calculate OEE depending upon the use of the measured values. See Ref. [3] for some of the most popular ways to calculate OEE for semiconductor industry.

22.7.3 Cost of Ownership

Availability and OEE are the most widely used high-level metrics, but they have the following shortcomings. They do not include:

1. Acquisition and operational cost
2. Effect of the production volume
3. Product scrap loss because of poor quality output
4. Consumable cost
5. Waste disposal cost
6. Taxes, insurance, and interest expenses.

To overcome the above shortcomings, SEMATECH developed a CoO model [4], which calculates the true CoO per good unit produced in a given time period, usually a calendar year.

The CoO depends upon the equipment acquisition cost, equipment reliability, equipment maintenance and operational costs, production throughput rate, throughput yield, and equipment utilization. The basic CoO is given by the following equation.

$$\text{CoO per unit} = \frac{\text{FC} + \text{OC} + \text{YLC}}{\text{P} \times \text{THP} \times \text{U} \times \text{Y}} \tag{22.20}$$

where

Fixed costs (FC). The FC are typically determined from a variety of items such as: purchase price, taxes and duties, transportation costs, installation cost, start-up cost, and training cost. It also depends upon allowable depreciation schedule and the length of the time period under consideration.

Operating costs (OC). Operating costs for a piece of equipment are consumable, material, maintenance and repair, parts, waste disposal, and operators for the time period under consideration.

Yield loss costs (YLC). Yield loss costs are those associated with lost production units that are directly attributable to equipment performance during the time period under consideration.

Time period (P). Time period under consideration, usually a calendar year expressed in hours.

Throughout rate (THP). Throughput rate is the actual average (for the time period under consideration) production rate of the equipment, expressed in parts per hour.

Utilization (U). As defined in Section 22.8.1, expressed in fraction.

Throughput yield (Y). Throughput yield also known as quality rate, is the fraction of good units produced. It is determined by

$$Y = \frac{\text{Total units produced} - \text{Defective units produced}}{\text{Total units produced}} \tag{22.21}$$

Table 22.3 contains an example of a simple CoO calculation. See [4] for more elaborate CoO calculations.

22.7.4 Hierarchy of Equipment Performance Metrics

Figure 22.3 depicts the hierarchy of equipment performance metrics. As shown in the figure, when we add time dimension to quality and safety, it becomes reliability. Reliability and maintainability jointly make up availability. When production speed efficiency and production defect rate are combined with availability, they become OEE. Acquisition and operational cost make up life cycle cost (LCC) See Ref. [1] for detailed description of LCC. When scrap, waste, consumables, tax, and insurance cost are added to LCC and the total is normalized by the production volume, it becomes CoO.

22.8 An Example of Reliability and High Level Performance Metrics Calculations

22.8.1 Given Values

Assume that the following values are given for the reliability and other performance metrics calculations.

Theoretical throughput = 50 wafer per hour
Productive time = 1200 h
Observed throughput = 40 wafer per hour
Number of defective wafers = 105
Number of failures = 3
Total unscheduled down time = 20 h (including supplier maintenance delay time = 5 h and repair time = 15 h)

TABLE 22.3 A Typical Simple Cost of Ownership (CoO) Calculations

CoO Input Data	
Equipment acquisition cost = $1,000,000	Equipment life = 5 years, straight line depreciation
Throughput rate = 20 units/h	Throughput yield = 0.98
Operation cost = $800,000/year in 2004	Part cost = $50,000/year in 2004
Labor rate = $50/h in 2004	Pre ventive Maintenance time = 10 h/month
Mean time between failure (MTBF) = 200 h	Mean time to repair (MTTR) = 2 h
Utilization = 75%	Inflation rate = 4% per year

CoO Calculations					
	Year				
Cost factors	2004	2005	2006	2007	2008
Depreciation ($)	200,000	200,000	200,000	200,000	200,000
Operational cost ($)	800,000	832,000	865,280	899,891	935,887
Repair and maintenance cost ($)	66,320	68,973	71,732	74,601	77,585
Yield loss ($)	250,000	260,000	270,400	281,216	292,465
Total cost ($)	1,116,320	1,160,973	1,207,412	1,255,708	1,305,937
Good unit produced	128,772	128,772	128,772	128,772	128,772
CoO per unit ($)	8.67	9.02	9.38	9.75	10.14

PM time = 50 h
Facility down time = 8 h
No operator standby time = 2 h
Non-schedule (holiday) time = 24 h.

22.8.2 Metrics Calculations

The following equipment performance metrics are calculated based on the above values.

Observed mean productive time between failures ($MTBF_p$) = (1200)/3 = 400 h
Observed mean wafers between failures (MWBF) = (1200 × 40)/3 = 16,000 wafers
Eighty percent lower confidence limit = 400 × 0.701 = 280.4 h for $MTBF_p$
Eighty percent lower confidence limit = 16,000 × 0.701 = 11,216 wafers for MWBF
Observed mean time to repair (MTTR) = (20 − 5)/3 = 5 h
Equipment uptime = 1200 + 2 = 1202 h
Equipment dependent down times (DT_E) = (20 − 5) + 50 = 65 h
Equipment dependent uptime (%) = (1202) × 100/(1202 + 65) = 94.87%
Supplier dependent uptime (%) = (1202) × 100/(1202 + 65 + 5) = 94.49%
Operational uptime (%) = (1202) × 100/(1202 + 65 + 5 + 8) = 93.90%
Total utilization (%) = (1200) × 100/(1202 + 65 + 5 + 8 + 24) = 92.02%
Simple OEE (%) = (1200 × 40 − 105) × 100/((1202 + 65 + 5 + 8 + 24) × 50) = 73.46%.

22.9 Four Steps to Better Equipment Reliability

Four basic steps to better equipment reliability are:

1. Know goals and requirements
2. Design-in reliability
3. Built-in reliability
4. Manage reliability growth during
   a. Reliability tests
   b. Field operations.

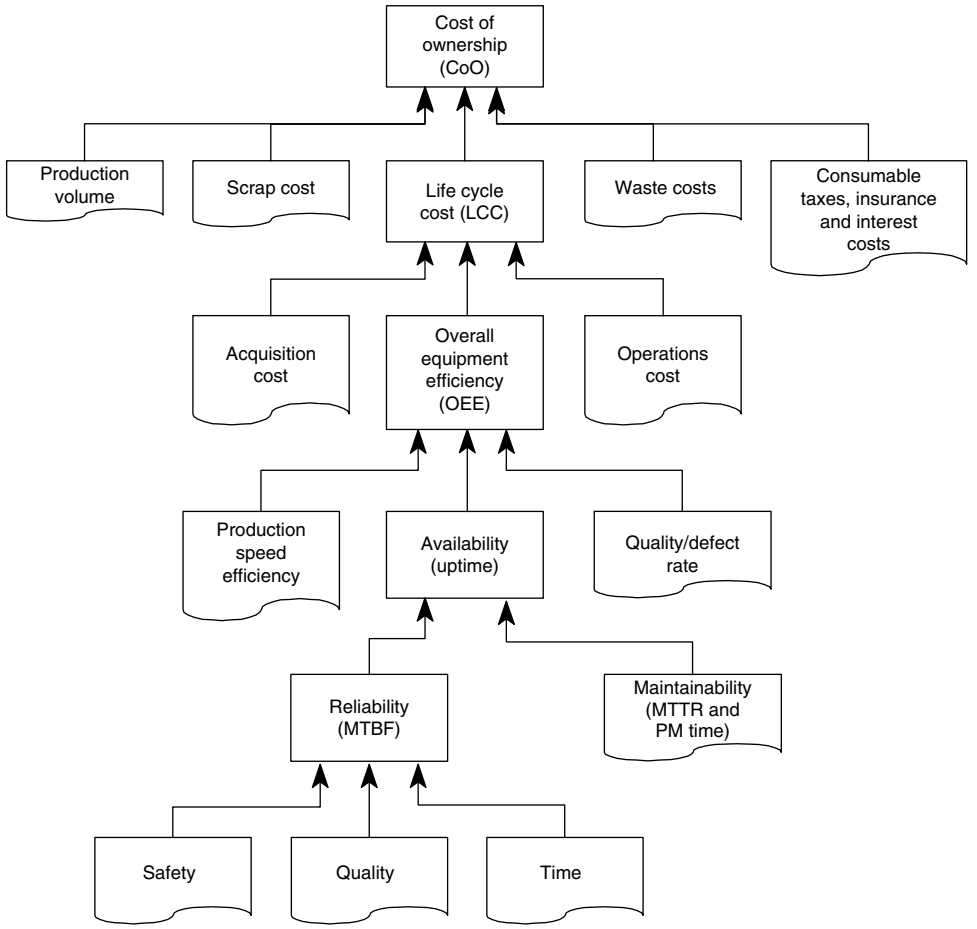


FIGURE 22.3 Hierarchy of high-level equipment performance metrics.

All reliability improvement activities belong to one of these steps.

22.9.1 Know Goals and Requirements

If what is required is unknown, then, it probably will not be achieved. Therefore, the first step to better reliability is to know the reliability goals and requirements, whether you are an equipment manufacturer or a user.

If you are an equipment manufacturer (supplier), you need to understand the exact reliability requirements of your customer

1. Be aware of the reliability level of your competitor's product
2. Know what reliability level is required in the market place.

Considering the above inputs, set reliability goals for the equipment line at the beginning of each equipment program. If you are a customer, then it is your responsibility to make sure that equipment suppliers know your exact requirements. In either case, the reliability goals or requirements must include, at minimum, the items shown in Table 22.4.

22.9.1.1 Goal Allocation

Once the system level goals are known, the equipment manufacturer must break down the equipment and system-level goals into "bite-size" goals for sub-systems, modules, and components. This makes it relatively easy for subsystem, module, or component engineers to achieve their respective product goals.

1. The process of breaking down the equipment and system-level goals into the next levels of sub-goals, based on some logical justification, is called apportionment, budgeting, or allocation. This process is just like breaking down division-level budgets into department-level budgets. One widely used method is known as Advisory Group on Reliability of Electronic Equipment (AGREE) allocation method. In this method, appropriate weight factors, based on the complexity and criticality of the components are used in the calculations as shown below.

$$\mathrm{MTBF}_i = \mathrm{MTBF}_s\Big(\sum_{i=1}^{h} W_i\Big)/W_i \tag{22.22}$$

where MTBF_i, goal MTBF of the ith component; MTBF_s, system level MTBF goal; W_i, weighting factor for ith component based on its complexity (1 = simple, 10 = most complex, sometime the weight factors are based on the inherent failure rate).

An example of AGREE method calculations

If a system consists of five modules, the system level MTBF_S goal = 500 h, and the weight factors for each module are as follows:

For Module 1, $W_1 = 6$ For Module 2, $W_2 = 10$ For Module 3, $W_3 = 3$, For Module 4, $W_4 = 6$ For Module 5, $W_5 = 5$

Then:

MTBF goal for Module 1 (MTBF_1) = 500 × (30/6) = 500 × 5 = 2500 h
MTBF goal for Module 2 (MTBF_2) = 500 × (30/10) = 500 × 3 = 1500 h
MTBF goal for Module 3 (MTBF_3) = 500 × (30/3) = 500 × 10 = 5000 h
MTBF goal or Module 4 (MTBF_4) = 500 × (30/6) = 500 × 5 = 2500 h
MTBF goal for Module 5 (MTBF_5) = 500 × (30/5) = 500 × 6 = 3000 h.

TABLE 22.4 Must Include Items in the Reliability Requirements/Goals

Item	Examples
Reliability metric and level that equipment should attain	Mean time between failure (MTBF) = 700 h
Time factor, age at which equipment should attain the reliability level	Four months after installation
Operational conditions:	
Temperature	Temperature range: 70°F–75°F
Humidity	Humidity range: 40%–45% Relative Humidity
Duty cycle	12 h/day
Throughput rate	15 wafers/h
Process to be used	High density plasma etch
Operator skill level	Grade 12 or equivalent
PM policies to be followed	Monthly PM policy
Shipping mode	Air-cushioned truck
Installation procedure	Install by a special installation team
Confidence level for attaining the reliability level	80% confidence that observed MTBF is equal or greater than the goal value
Acceptable evidences for attaining the required reliability level	Attaining the reliability level based on the field data, four months after installation

22.9.2 Design-In Reliability

This is the most important and elaborate step to achieve better equipment reliability. Design-in reliability is a process in which reliability improvement goals are considered concurrently with other technical aspects at every activity of the design phase. Figure 22.4 depicts the design-in reliability process. Six major blocks of the process are

1. Use proper parts properly
2. Use proper design techniques
3. Design to withstand effect of external factors
4. Avoid failures through scheduled maintenance

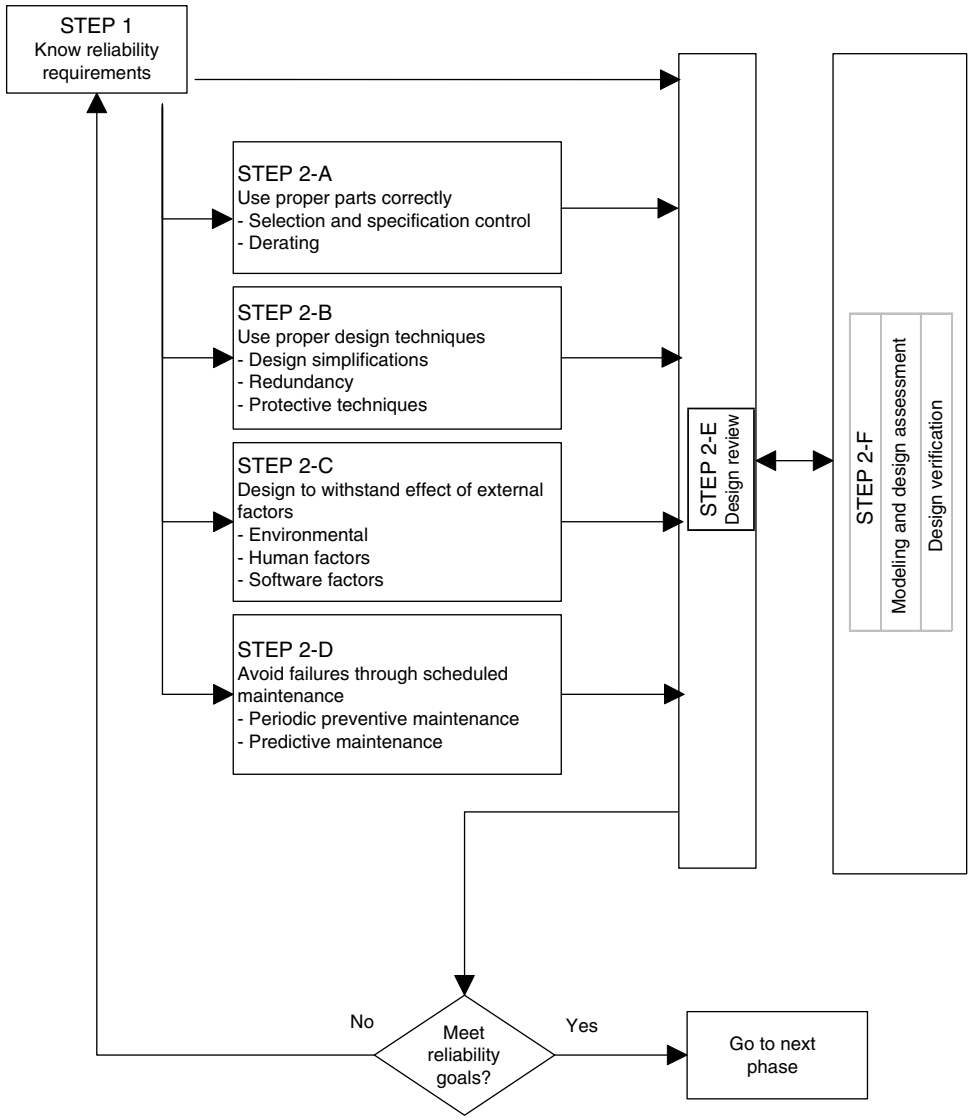


FIGURE 22.4 Process of design-in reliability.

5. Hold design reviews
6. Assess reliability of the design using modeling techniques.

Reference [5] contains a detailed description of each block. The following is a brief summary.

22.9.2.1 Use Proper Parts Properly

Use of the proper parts properly makes up the most crucial block of the design-in reliability process. It consists of the following activities:

Part selection. Before selecting any part and its supplier, determine the part type needed to perform the required functions and the environment in which it is expected to operate. The general rule for part selection is that, whenever possible, the designer should strive to use proven parts in the design and select a supplier who has proved historically to meet or exceed the part reliability requirements.

Part specification. For each reliability sensitive part, its procurement specification should include:

1. Details of the intended application(s)
2. Reliability requirement level for the intended application(s)
3. Part screening procedure
4. Part qualification procedure
5. Acceptable evidences for attaining the required reliability level

Part derating. Once the part is selected, perform an analysis to compare the expected stress level for the intended applications with those of the part's rated (capacity) stress level. A technique known as *derating* is used to improve design reliability. In this technique, a part is selected, so that it will operate at less severe stress than it is rated as capable of operating. For example, if the expected power level is 10 W for a device, select parts that are rated for significantly higher than 10 W power. Use the appropriate derating factors for various electronics components given in Table 22.5.

22.9.2.2 Use Proper Design Techniques

Use of proper design techniques is another most crucial block of the design-in reliability process. It consists of the following activities:

Design simplification. Anything that can be done to reduce the complexity of the design will, as a general rule, improve reliability. If a part is not required, eliminate it from the design. Wherever possible, reduce the number of parts through combining functions.

Redundancy. This is one of the most popular methods in design to achieve the needed level of reliability. Redundancy is the provision of using more than one part to accomplishing a given function so that all parts must fail before causing a system failure. Redundancy, therefore, permits a system to operate even though some parts have failed, thus increasing system reliability.

For example, if we have a simple system consisting of two identical redundant (parallel) parts, the system MTBF will be 1.5 times that of the individual part MTBF.

Protective Technique. This design technique includes a means to prevent a failed part or malfunction from causing further damage to other parts. The following are some of the popular protective techniques used in equipment designs:

1. Fuses or circuit breakers to sense excessive current drain and to cut off power to prevent further damage
2. Thermostats to sense over-temperature conditions and shut down the part or system operation until the temperature returns to normal
3. Mechanical stops to prevent mechanical parts from traveling beyond their limits
4. Pressure regulators and accumulators to prevents pressure surges

TABLE 22.5 Derating Factors for Commonly Used Components

Component	Stress Category	Derating Factor
Capacitors, general	Voltage	0.5
Capacitors, ceramic	Voltage	0.5 at $<85°C$ 0.3 at $<125°C$
Capacitors, supermetallized, plastic film, any tantalum	Voltage temperature	0.5 at $<85°C$ Less than 85°C
Capacitors, glass dielectric, fixed mica	Voltage temperature	0.5 at $<85°C$ Less than 85°C
Connectors	Current	0.5
	Voltage	0.5
	Temperature	Less than 125°C
Quartz crystals	Power	0.25
Diodes	Voltage	0.75
	Current	0.5
EMI and RFI filters	Voltage	0.5
	Current	0.75
Fuses	Current	0.7 at $<25°C$
		0.5 at $>125°C$
Integrated circuits (all kinds)	Voltage	0.7
	Current	0.8
	Power	0.75
Resistors (all kinds)	Voltage	0.8
	Power	0.5
Thermistors	Power	0.5
Relays and switches	Current	0.75 for resistive load
		0.4 for inductive load
		0.2 for motors
		0.1 for filament
Transistors	Breakdown voltage	0.75
	Junction temperature	Less than 105°C
Wires and cables	Current	0.6

5. Self-checking circuits (and software) to sense abnormal conditions and make necessary adjustments/compensations to restore normal conditions
6. Interlock to prevent inadvertent operations
7. Homing sequence for computer shut-downs

22.9.2.3 Design to Withstand Effect of External Factors

The operating environment is neither forgiving nor understanding. It methodically surrounds and affects every part of a system. If a part cannot sustain the effects of its environment, then reliability suffers. First, the equipment manufacturer must understand the operating environment and its potential effects. Then, he must select designs and materials that are able to withstand these effects or he must provide methods to alter and control environmental conditions within acceptable limits.

Equipment design engineers must consider means to withstand the following external factors affecting reliability:

1. Heat generation causing high temperatures
2. Shock and vibration
3. Moisture
4. High vacuum
5. Explosion
6. Electromagnetic compatibility
7. Human use
8. Utilities supply abnormalities (power, water, gases, chemicals, etc.)
9. Software design.

22.9.2.4 Avoid Failures through Scheduled Maintenance

One way to improve reliability is to minimize the number of failures that occur during operation. This can be achieved in two ways:

1. Select parts that fail less frequently
2. Replace a part before its expected failure time.

The latter method is known as *scheduled maintenance* (SM). This technique is used when it is not feasible to find a part that fails less frequently. If such a situation is properly comprehended during the design phase, it can be avoided through one of the following SM techniques.

Periodic preventive maintenance. This is a fixed-period-driven maintenance procedure in which parts that are partially worn out, aged, out-of-adjustment, or contaminated, are replaced, adjusted, or cleaned before they are expected to fail. This way, system failures are forestalled during the system operations thus reducing the average failure rate.

Predictive maintenance. This is a condition-driven scheduled preventive maintenance program. Instead of relying on a fixed period of life units to schedule maintenance activities, predictive maintenance uses direct monitoring of appropriate indicators to determine the proper time to perform the required maintenance activities.

22.9.2.5 Design Review

Design reviews are an essential element of the design-in reliability process. The main purposes of a design review are to assure that:

1. Customer requirements are satisfied
2. The design has been studied to identify possible problems
3. Alternatives have been considered before selecting a design
4. All necessary improvements are based on cost trade-off studies.

Conduct design reviews on a regular basis from the initial design feasibility study through the pilot production phase. An effective design review team should have representation from each functional area involved in developing the equipment.

22.9.2.6 Reliability Assessment of the Design

Once the equipment design starts taking shape, it must be assessed to determine its reliability level and the system level effect of failure rate of each part. Modeling techniques determine analytical value of the reliability level. Failure mode and effects analysis (FMEA) technique determines the system level effect of part failures. Failure mode and effects analysis also identifies improvement opportunities. There are many commercially available software packages that perform modeling and FMEA.

22.9.2.7 Design Verification Tests

Once design is firmed, conduct design verification test to determine the reliability level under the expected use conditions (see Section 22.10). If the observed reliability level is lower than the expected, identify areas for improvement and implement corrective actions (design changes).

22.9.3 Build-In Reliability

Building-in reliability is a process that assures all parts, subsystems, modules that are made and assembled according to engineering drawings and specifications without degrading the designed reliability or introducing new failure modes. Important steps of this process are:

Assembly instructions. Prepare detailed instructions for each assembly step. These instructions should include proper parts, materials, step-by-step assembly procedures, tools, limitations, inspection procedures, etc.

Training. To minimize assembly errors, it is essential that every assembly operator be trained in basic assembly methods and in all the assembly operations assigned to him or her

Burned-in. All parts and the system itself should be properly and adequately burned-in, debugged, or stress-screened before shipment.

Product reliability acceptance test (PRAT). Conduct a PRAT on randomly selected units before shipping to assure the reliability level of the product line as it is shipped.

Packaging and shipping. Equipment must be packed properly for the intended shipping mode. Select shipping mode that does not impart any undue stress on the equipment.

22.9.4 Manage Reliability Growth

Effectively managing reliability growth opportunities is a continuous improvement process. Equipment manufacturers learn from in-house reliability tests (see Section 22.10) or actual field experience. All problems observed during the reliability tests (either in-house or customers place) are documented and given to a central body (such as Failure Review Board (FRB)) for further analysis and disposition. If required, corrective actions are developed and implemented.

Similarly, during field operations, customer works with the equipment supplier to collect, record, and analyze equipment failures, both hardware and software. They capture predetermined types of data about all problems observed with a particular equipment line and submit the data to the supplier. The FRB, at the supplier's site, analyzes the failures. The resulting analysis identifies corrective actions that should be developed, verified, and implemented to prevent failures from recurring.

A popular system named "Failure Reporting and Corrective Action System" (FRACAS) is used to manage this process. As shown in Figure 22.5, FRACAS is a closed-loop feedback communication channel to report, analyze, and remove failure causes.

Now let us look at three key elements of FRACAS in more detail.

22.9.4.1 Failure Data Reporting

All the failures, observed either during in-house test or at customer's site, must be recorded so all relevant and necessary data is captured in a systematic manner. A simple, easy-to-use form that is tailored to the respective equipment line should be used to record and report failure data. (Figure 22.6 depicts a typical failure reporting form). If the data volume justifies the cost of administering FRACAS, the data form can be computerized to communicate failure data. Internet and electronic mail are the most recent ways to report failure data.

22.9.4.2 Failure Review Board

The FRB is a multifunctional self-managed team that reviews, facilitates, and administers failure analysis. It also participates in assigning, developing, verifying, and implementing the resulting corrective actions. To do this job effectively, all the functional departments involved in the product line must participate on the FRB. Also FRB members must be empowered to assume responsibility, investigate failure cause, develop corrective actions, and ensure implementation of corrective actions.

22.9.4.3 Corrective Action

Any systematic action taken to eliminate or reduce the frequency of equipment failure (hardware or software) is a corrective action. Such actions may include part designs or material changes, part supplier changes, assembly procedure changes, maintenance procedure changes, operational changes, training changes, or software changes.

All corrective action plans and their verification and implementation should be reviewed by the FRB on a regular basis. The FRB should also maintain a log of the corrective action status including open and closed corrective actions.

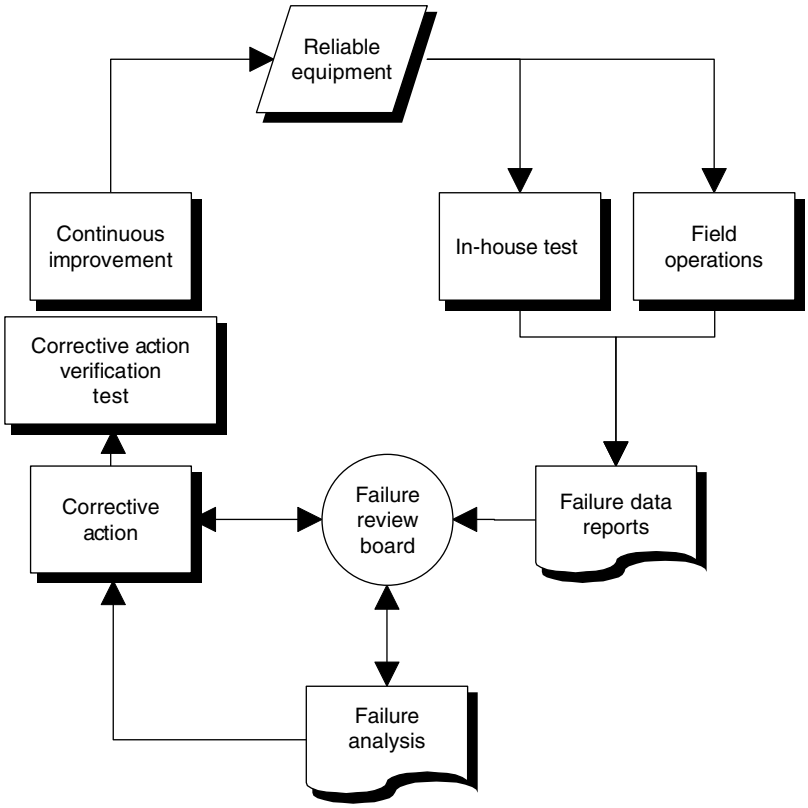


FIGURE 22.5 Failure Reporting and Corrective Action System (FRACAS) process flow.

22.10 Reliability Testing

No matter how many or how extensive the analyses we perform to calculate the reliability level of equipment, it is almost impossible to calculate the effect of all the factors that affect the reliability level. Even after engaging a meticulous reliability modeling software programs to do the reliability level calculations, we cannot theoretically derive the exact reliability level that will be observed in a reliability test or when equipment is installed at the customer's site. This inability of our theoretical efforts necessitates performing reliability tests to find the actual reliability level of the equipment configuration of interest. The tests validate theoretical calculations and provide a proof for the expected performance indices.

Reliability testing is also a very important activity of the reliability improvement programs. Information generated during the reliability test is vital to design engineers for initial designs and subsequent redesigns or refinements, and to the manufacturing engineers for fine-tuning the manufacturing process. The reliability tests also provide vital information to program managers showing technical progress and problems of an equipment line.

Reliability tests can be performed at any level of integration, i.e., at the component level, part level, module level, subsystem level, or system level. Not only that, they can be performed during any equipment program life cycle phase.

Among the numerous reasons to conduct reliability tests are:

1. To determine the reliability level under the expected use conditions
2. To qualify that the equipment line meets or exceeds the required reliability level

SERVICE REPORT NO. _____

Customer	Customer System ID	Module	Received By:	
Reported By	Equipment Serial No.	Module Serial No.	Ref. Report NO.	
Phone No.	Contact		Date	Time
Service Category: ☐ Install ☐ Courtesy ☐ Warranty ☐ PM ☐ Development ☐ Training ☐ Telephone Fix ☐ Update ☐ UM ☐ Others				
Problem/Symptoms/Reason for Service Call				
Repair & Maintenance Action(s)				
Problem Cause				

Part Name	Part Number	Replaced	Cleaned	Comments

Status of Equipment Upon Leaving	Start Date	Time
	Complete Date	Time
Customer Remarks	Service Engineer's Signature	
	Customer's Signature	

FIGURE 22.6 A typical failure report form.

3. To ensure that the desired level is maintained throughout the equipment life cycle phases.
4. To improve reliability by identifying and removing root failure causes.

22.10.1 Types of Reliability Tests

Since reliability testing is included in all the equipment life cycle phases, and they are conducted for numerous reasons, it follows that the testing includes many types of tests. The following reliability tests are commonly seen during a typical equipment program.

1. Burn-in tests
2. Environmental stress screening (ESS) tests

3. Design verification tests
4. Reliability development/growth tests
5. Reliability qualification tests.

Burn-in Tests. These test are conducted to screen out parts that fail during the early life period. They are performed at part, subsystem, or system level. Most failures observed during these tests are due to manufacturing workmanship errors, poor quality parts, and shipping damage. System level burn-in tests are also known as debug tests.

Environmental Stress Screening Tests. As the title indicates, the ESS tests are conducted in an operating environment that is harsher (stressed) than the normal environment of expected use. The main purpose of the test is to weed out parts that, otherwise, would not fail under normal operating environment. This test increases confidence that all received parts are of good quality and they will last longer (i.e., have better reliability).

Design Verification Tests. Design verification tests are conducted to ensure that a desired reliability level is achieved during the design phase under the expected use conditions. Most of these tests are run at the system level.

Reliability Development/Growth Tests. Reliability development/growth tests are conducted to ensure that a desired reliability level is achieved during a given equipment program life cycle phase and it is improving (growing) as the program moves further along the life cycle phases. Most of these tests are run at the system level.

Reliability Qualification Tests. These test are conducted to qualify that the component or equipment meets or exceeds the reliability level. These are pass–fail tests. If the demonstrated reliability level is equal or better than the required level and confidence, the equipment (or its program) is considered as meeting the requirement–passing the test or qualifying the equipment.

22.10.2 Generic Steps for Reliability Tests

Three overall steps of any typical reliability test are:

1. Test plan development
2. Test conducting
3. Test data analysis and reporting.

22.10.2.1 Test Plan Development

A well thought-out reliability test plan includes the following:

1. Test objectives
2. Hardware, software, and process to be used
3. Operational stresses and environment
4. Resources required (including consumable)
5. Sample size and test length
6. Test procedure
7. Data to be acquired
8. Data form to be used
9. Data analysis techniques
10. Data reporting and reviewing procedures
11. Pass–fail criteria, if required
12. Expected outcome for each test
13. Types of test reports
14. Schedule of key test activities.

The reliability test plan should be formally documented and approved by high-level managers.

22.10.2.1.1 Test Length

The test length depends upon the desired confidence in the test results and the expected level of reliability (MTBF). Tests need to run long enough to increase confidence in the test results. However, we never have enough resources or time to test for an extended period. Therefore, statisticians have developed a method to determine the minimum test length needed to make correct decision with the required confidence in that decision. For repairable equipment, minimum test lengths are calculated to ascertain certain minimum MTBF level (target MTBF) with certain confidence, by

$$\text{Minimum Test Length with } P\% \text{ Confidence} = (\text{Target MTBF}) \times \omega \tag{22.23}$$

where P%, desired confidence level; target MTBF, MTBF to be proved or expected; ω, appropriate multiplier for P% confidence from Table 22.6.

For example, if we need to prove target MTBF of 100 h, with 80% confidence in the decision, the minimum test length is calculated as follows.

Target MTBF = 100 h
ω = 1.61 from Table 22.6 for 80% confidence level
These give a minimum test length = 100 × 1.61 = 161 h

If we observe no or one failure in the 161 h-long test, we meet the target MTBF of 100 h.

22.10.2.1.2 Test Conducting

During this step, the reliability test is conducted according to the test plan. All deviations from the formal test plan should be recorded and approved. A formal log of test events is kept to record key test parameters associated with each event.

22.10.2.1.3 Test Data Analysis and Reporting

All the data collected during the test are appropriately analyzed, and conclusions are made. Use the formulae given in Section 22.2, Section 22.4, Section 22.6, and Section 22.7 to determine the observed reliability level, the associated confidence limits, and other performance parameters. You can also use SEMI E10 [2] formula to calculate the desired metrics. Failure Review Board and other interested groups should review the test data, results, and conclusions. To close a test project, a formal test report must be issued containing test objectives, test procedures, findings, conclusions, and recommendations.

22.10.3 Reliability Tests throughout the Equipment Program Life Cycle Phases

Reliability tests are scattered throughout the equipment program life cycle phases. They play a very important part in the reliability improvement process. Table 22.7 lists the appropriate tests for each phase.

TABLE 22.6 Minimum Test Length Multiplier ω

	Confidence P%						
	10	20	50	75	80	90	95
Multiplier ω	0.11	0.22	0.69	1.38	1.61	2.30	2.99

TABLE 22.7 Reliability Tests throughout the Equipment Life Cycle Phases

Life Cycle Phase	Reliability Test
Concept and feasibility	No formal reliability test
Design	Part-level reliability qualification
	Design verification and reliability development
	Accelerated test
Prototype	Part-level reliability qualification
	Design verification
	Reliability qualification
	Reliability growth
	Accelerated test
Pilot production	Burn-in
	Environmental stress screening
	System-level reliability qualification
	Accelerated test
	Reliability growth
Production	Burn-in
	Environmental stress screening
	Reliability qualification
	Product reliability acceptance test
	Accelerated test
	Reliability growth
Phase out	None recommended

22.11 Use of Equipment Reliability Discipline in Business Practices

To achieve equipment reliability goals/requirements, equipment manufacturers and users must include equipment reliability in their business practices as described below.

Equipment suppliers. They must use equipment reliability throughout the various equipment-life cycle phases as shown in Table 22.8. Beside these activities, equipment suppliers must insists that their component suppliers use equipment reliability in their business practices. Similarly, all dealing with their customers must include equipment reliability. For example, any reference to reliability requirement in either request for quotation (RFQ) or purchase order (PO) documents must

TABLE 22.8 Uses of Equipment Reliability in Suppliers' Business Practices

Life Cycle Phase	Activities Using Equipment Reliability Discipline
Concept and feasibility	Goal setting
	Apportionment
	Reliability plans
	Preliminary modeling
Design phase	Design-in reliability
	Design assessment and modeling
	Design verification testing
	Failure mode and effect analysis (FMEA)
	Part life tests
	Design review
Prototype	System test and reliability level assessment
Phase and	Design review
Pilot production phase	Failure reporting, analysis, and corrective action system (FRACAS)
Production phase	FRACAS Product reliability assurance tests (PRAT)

include items shown in Table 22.4. If a customer does not specify reliability requirement clearly, the supplier must let the customer know that specific reliability requirements are lacking in the PO and must be clarified.

Equipment users. A need for equipment reliability discipline begins when a user initiates an equipment search. At that time, the user must know the reliability requirements of the equipment they are going to acquire. These requirements must include items shown in Table 22.4. Whenever a user decides to evaluate equipment in his own factory, the reliability level calculations must be based on the formula given in Section 22.2, Section 22.4, Section 22.6, and Section 22.7. All the RFQ's and PO's must include appropriate reliability metrics. After the equipment has been purchased and installed in the factory, equipment users must accurately track the reliability performance level. They must implement continuous equipment improvement activities, such as failure reduction or OEE improvement programs. Table 22.9 summarizes the uses of equipment reliability by equipment users in their business practices.

22.12 SEMI E10

This chapter cannot end without describing SEMI E10, a Semiconductor Equipment and Materials International (SEMI) specification for definition and measurement of equipment RAM. A task force (consisting of semiconductor manufacturing equipment suppliers and users) developed it under the SEMI Metrics Committee. The SEMI E10 Guideline was issued in 1986 and revised several times. With a revision in 1996, SEMI E10 became a SEMI Standard and SEMI specification in 2000. See Ref. [2] for the latest revision of SEMI E10 (SEMI E10-0304^{E}).

22.12.1 Benefits of Using SEMI E10

To create synergy between semiconductor equipment suppliers and users, they must work together for mutual gains, understand RAM expectations of each other, and speak the same language when talking about equipment RAM metrics. SEMI E10 Specification provides this language (a common basis for communication between users and suppliers of semiconductor manufacturing equipment) by providing specifications for measuring RAM metrics of equipment in manufacturing environments. For equipment suppliers, SEMI E10 RAM metrics are useful at each product life cycle phase, from early equipment design and development through production. From equipment users point of view, SEMI E10 provides an industry-wide and company-wide uniform specification to collect, analyze, track, compare (machine to machine, wafer fab to wafer fab, and industry wide), and report equipment RAM data. Accurate data collection of time allocation in each state is essential to calculate the accurate RAM metrics. Automation efforts to collect the data are also based on SEMI E10 definitions and formulae. Both CoO and factory capacity analyses use SEMI E10 RAM metrics. They also provide a basis for specifying reliability performance in equipment PO agreements. The long-term benefits of SEMI E10's

TABLE 22.9 Uses of Equipment Reliability in Equipment Users' Business Practices

Users Activity	Activities Using Equipment Reliability Discipline
Equipment search	Reliability specifications and observed value calculations
Equipment evaluations	Observed failure rate (or MTBF) and confidence limit calculations
Sending RFQ's	Reliability specifications and all the reliability metrics used in the RFQ's
Sending purchase orders	Reliability specifications and all the reliability metrics used in the PO's
Equipment performance tracking	Observed failure rate (or MTBF) and confidence limit calculations
Equipment improvement programs	Observed failure rate (or MTBF) and confidence limit calculations
Using high-level equipment performance metrics	Reliability elements of the high-level metrics

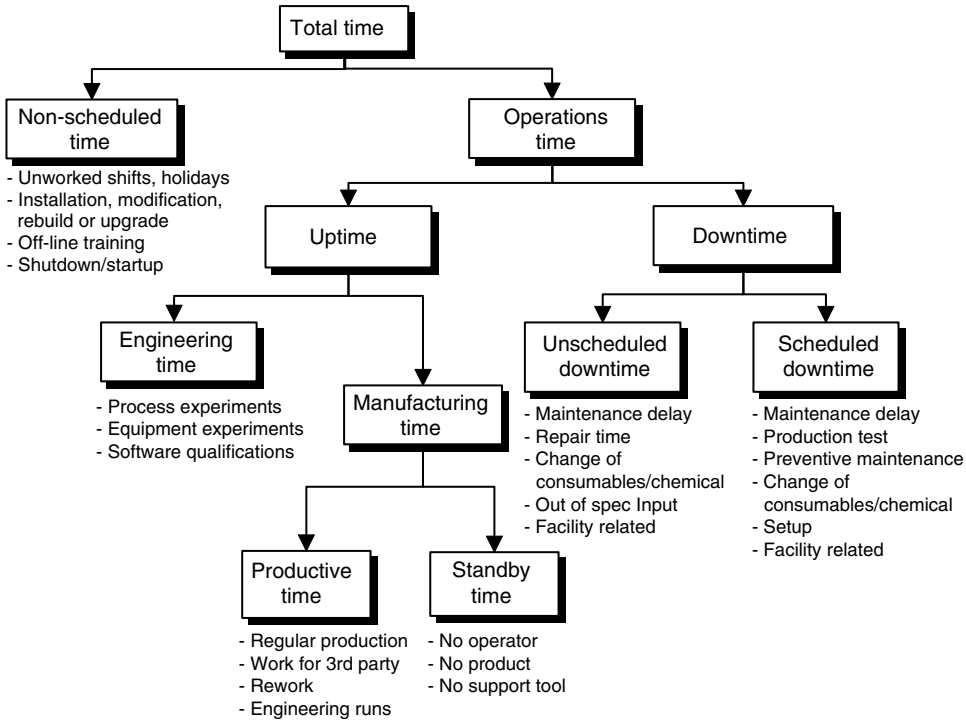


FIGURE 22.7 Categorization of total time into SEMI E10 states. (SEMI E10-0304^{E}. Specification for Definition and Measurement of Equipment Reliability, Availability, and Maintainability (RAM), SEMI International Standard, Equipment Automation/Hardware, San Jose, CA 2004.)

international acceptance and use are the improved relationships between users and suppliers of semiconductor (SC) manufacturing equipment that will stimulate a spirit of cooperation and partnership, promoting further improvements in equipment performance. These will lead to greater business success for both users and suppliers.

22.12.2 Key Elements of SEMI E10

Two key elements of SEMI E10 are (a) events, scheduled, unscheduled, or non-scheduled, that stop equipment from performing its intended functions, and (b) arrival and departure times for the events.

Events are categorized as scheduled or unscheduled. Unscheduled events are called failures. SEMI E10-0304^{E} defines the failure events as follows.

Any unscheduled downtime event changes the equipment to a condition, where it cannot perform its intended function. Any part failure, software or process recipe problem, facility or utility supply malfunction, or human error could cause the failure.

SEMI E10-0304^{E} uses the arrival and departure times of the scheduled, unscheduled, or non-schedule events to break down total calendar time into various time blocks as shown in Figure 22.7. These time blocks are defined as equipment states and become the basis for equipment RAM metric calculations similar to those given in Section 22.2, Section 22.4, Section 22.7, and Section 22.8 [2].

References

1. Dhudshia, V. H. *Hi-Tech Equipment Reliability: A Practical Guide for Engineers and the Engineering Managers*. Sunnyvale, CA: Lanchester Press, 1995.
2. SEMI E10-0304[E]. *Specification for Definition and Measurement of Equipment Reliability, Availability, and Maintainability (RAM)*. San Jose, CA: SEMI International Standard, Equipment Automation/Hardware, 2004.
3. SEMI E79-0299. *Standard for Definition and Measurement of Equipment Productivity.* San Jose, CA: SEMI International Standard, Equipment Automation/Hardware, 1999.
4. SEMATECH. *Cost of Ownership Model.* Austin, TX: SEMATECH, Inc., 1992 (Technology Transfer #91020473B-GEN).
5. SEMATECH. *Design Practices For Higher Equipment Reliability—Guidebook*. Austin, TX: SEMATCH, Inc., 1993 (Technology Transfer #93041608A-GEN).

23

Overview of Process Control

23.1 Introduction to Control of Systematic Yield Loss **23**-2

23.2 The Control-Type Categories **23**-4
Abnormality Control Methods • Compensation (Target Tracking) Control Methods • Advanced Process Control: Combination of Both

23.3 History of Process Control in Semiconductor Manufacturing **23**-7

23.4 Characterization of Control Needs in Semiconductor Manufacturing **23**-9
The Four Expected Sources of Variation Requiring Compensation • Timescale of Variations • Pilots, Look-Aheads, Metrology, and Operational Practices

23.5 Basic Concepts of All Control Techniques **23**-14
Process Qualification • Process Capability Indices • Types of Errors: False Positives and False Negatives

23.6 Specific Abnormality Detection and Control Methods **23**-18
Univariate Statistical Process Control • Fault Detection by Testing That There Is a Fault • Other Abnormality Control Methods and Use of Equipment Signals • Definition of the Sensitivity vs. Robustness Challenge

23.7 Specific Compensation Control Methods **23**-20
When Are Benefits Realizable? • Controller Goals: Tracking the Target, Rejecting Disturbances, and Ignoring Noise • Feedback/Feedforward Control • Common Compensation Control Methods Used for Run-to-Run Control • Real-Time Compensation Control Methods

23.8 Monitoring the Supervisory Run-to-Run Controller and the Controller System Advanced Process Control **23**-34
The Other Type I, Type II Errors: Detection of Change in Overall System • Methods for Monitoring the Supervisory Controller

23.9 Continuous Process Improvement **23**-35
Benefits of Reducing the *Effective* Noise of the System • Comparing How Different Machines React

23.10 Summary **23**-36

23.11 Acronyms and Glossary **23**-36

References **23**-37

Further Reading **23**-40

Stephanie Watts Butler
Texas Instruments, Inc.

23.1 Introduction to Control of Systematic Yield Loss

There are several different chapters in this handbook dedicated to control. As these chapters demonstrate, there are several types of control, each aimed at removing yield loss due to a particular source. All of the various control methods combined in their entirety can be viewed as "factory control." Other articles have been focused on this synergistic concept of combining the control techniques of different areas and data sources into one holistic approach to controlling the fabrication facility [1]. Thus, this chapter will not discuss the concept of factory control further. However, the readers should consider this concept as they learn more of each of the control methods in each chapter.

The increasing importance of systematic yield loss has been documented by a study by Keithley Instruments: "among the findings were that wafer misprocessing new accounts for more yield and reliability problems than contamination" [2]. This misprocessing includes both unintentional and intentional sources. Unintentional misprocessing includes operator and automation errors. Intentional misprocessing is the use of a recipe as requested by the engineer, but which recipe, coupled with the current state of the equipment, will result in less than desired results. This chapter will focus on methods other than those for defect and contamination control. In other words, this chapter will focus on methods for detecting and controlling misprocessing, also termed systematic yield loss. Consequently, for the remainder of the chapter, the term "process control" will be used to designate controlling the equipment and the process to ensure that desired results are achieved. Another reason for the application of process control is the need for improved productivity. In order to meet the historical trend of 30% per function per year reduction in cost, fab and equipment productivity must improve. In some cases, the yield is already high, but the overhead to achieve that yield is unacceptably costly. As this chapter demonstrate, application of certain types of control lead to improved productivity due to the reduction in pilot and look-ahead usage, with the yield remaining the same or even improving.

Systematic yield loss and misprocessing can be considered to be from two sources. The first source is an abnormality, i.e. unusual process behavior. The second source is expected, but undesirable, variation. Thus, control methods can be divided into three categories:

- Methods based upon detecting abnormalities and correcting them
- Methods based upon actively compensating for expected sources of variation
- Methods based upon a combination of compensation and abnormality detection

In practice, which control method is used is based upon how the engineer and organization view the situation. In other words, two people may view the same situation differently and consequently feel differently about which method is more appropriate. The purpose of this chapter is to provide the readers sufficient knowledge of the different methods and their assumptions so that they can judge which method is most appropriate. In addition, this chapter will provide a cursory overview of some of the emerging techniques so that the readers can identify new technology they feel is necessary for the future. How to locate the additional knowledge, information, and resources required for implementation of any of the methods is also provided.

Figure 23.1 explains what and how information is provided in this chapter, and the corresponding section number. Interdependencies of the various sections are shown so that the reader can decide which sections they may desire to read first. The chapter begins with a more complete definition of the three control categories listed above. All control methods discussed in this chapter will be relegated to one of the three control-type categories. A history of how control developed in the semiconductor industry enables the reader to understand why certain techniques have become popular and how different supplies came into being. To comprehend what control techniques are appropriate, a thorough understanding of the characteristics of the control need is necessary, i.e., the ability to dissect any systematic yield loss into its control characteristics is necessary before one can judge the appropriateness of any technique. The dissection of the control need characteristics also identifies the behaviors of the abnormal and expected

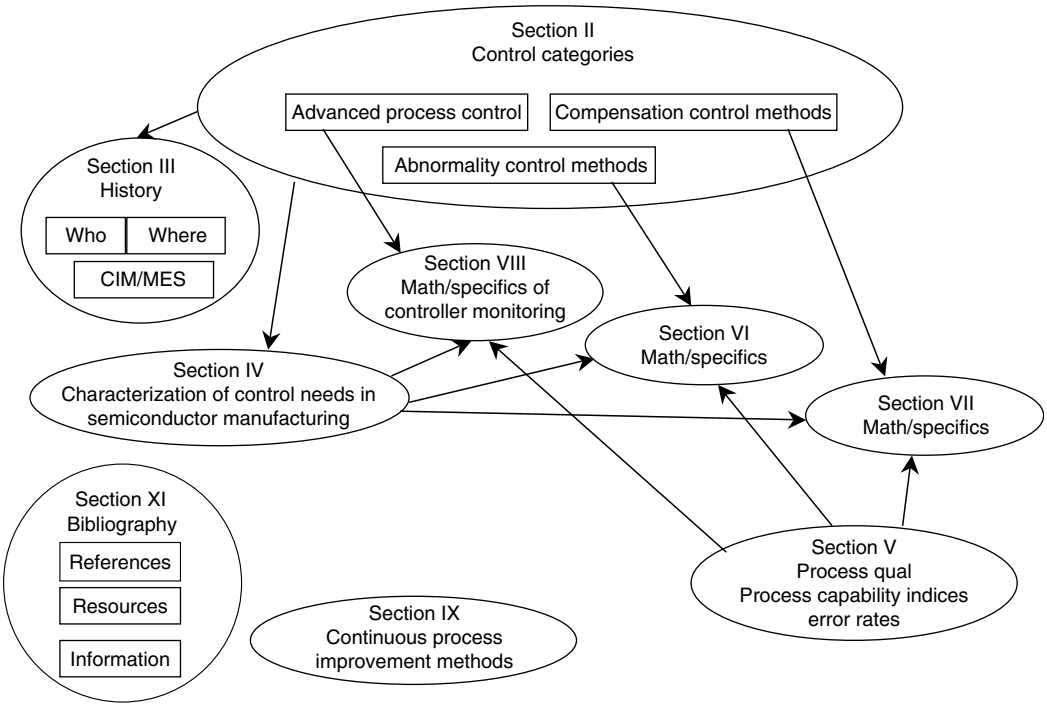


FIGURE 23.1 Explanation of what is covered in which sections.

variation and the situation in which it will occur. This explanation of the sources of systematic yield loss and how it relates to the different categories of control methods is the foundation for the entire chapter.

The chapter then switches to preparation for the introduction of specific technologies, with first the explanation of process qualification and process capabilities and their role in control. The chapter then becomes more technical as a review of some of the mathematics highlights the assumptions involved. Because of the importance of the concept of error rates which are encountered explicitly in abnormality control methods and implicitly in compensation control methods, a simplified discussion of error rate cause and relationships will be given before any specific techniques are presented. This review of errors will be referred to frequently in the subsequent sections, since optimizing error rates is one of the predominant drivers when determining which control technique to use. Then specific abnormality and compensation control techniques are presented. The compensation control section also stresses the requirements for successful application of compensation control, i.e., just because the systematic loss is high does not always mean one can apply compensation control to reduce it. The mathematics of abnormality control methods is not covered extensively because there are many excellent texts and articles, which cover this area. Instead, mathematical concepts that are then revisited in the section on compensation control are introduced.

The chapter concludes with a short review of the final, but most important, items of any control system, i.e., monitoring the supervisory controller and continuous process improvement. A bibliography is also included which the reader should find invaluable. Due to the large volume of literature, it would be impossible to review each article. Instead, the bibliography allows the reader to find further information about specific techniques and applications. The bibliography includes pointers to companies and web pages as well. A glossary provided at the end of the chapter assist the reader with definitions and acronyms. It is hoped that the material presented in this chapter will provide a strong foundation to allow the reader to comprehend other materials referred to in the bibliography.

23.2 The Control-Type Categories

As noted above, systematic yield loss and misprocessing can be considered to be from two sources. The first source is an abnormality, i.e., unusual process behavior. The second source is expected, but undesirable, variation. Thus, control methods can be divided into three categories:

- Methods based upon detecting abnormalities and correcting them
- Methods based upon actively compensating for expected sources of variation
- Methods based upon a combination of compensation and abnormality detection

Each of the above categories described below. The major supposition of each of the methods will be elaborated. The characteristics of the abnormal and expected variation and the situations in which it will occur described in the Section 23.4 characterization of control needs in Semiconductor manufacturing.

23.2.1 Abnormality Control Methods

Table 23.1 lists various groups of abnormality control method. The supposition is that there is normal variation ("common cause") and abnormal variation ("special cause"). Compensation is assumed to be undesirable or impossible for normal variation, while abnormal variation must be fixed, usually by repair performed by a human. Thus, these techniques can also be considered manual control. The groups of methods listed in Table 23.1 vary due to the source of data used by the method, but also by the final purpose of the method. Thus, note that some of the methods do not actually include a methodology for what action to take upon detecting an abnormality, i.e., they are really monitors only rather than controllers.

The basis of all the methods is the use of a technique to detect the abnormal variation (fault) in the presence of normal variation. Within a group of methods, as well as between group to group the biggest difference is the particular fault detection technique used. Because of the normal variation, one encounters difficulty in developing a technique that *always* detects abnormal variation without falsely identifying some normal variation as a fault also. While different fault detection techniques are developed to be more mathematically appropriate to the physics of the situation, they are also developed to achieve better error rates of erroneously declaring an abnormality. Fortuitously, techniques that are more appropriate mathematically for the situation usually yield better error rates. The mathematics of the error rates will be presented later.

23.2.2 Compensation (Target Tracking) Control Methods

Table 23.2 lists various groups and alternative terms of compensation control methods; Run-to-Run control is also known by a variety of terms that are listed in Table 23.3. The supposition in compensation control is that there is expected non-random variation and random noise. Compensation is assumed desirable and possible for the expected non-random variation. Classically in the other industries in which

TABLE 23.1 Terms and Groupings of Process Control Methods That Are Based upon Abnormality Detection and Correction (Manual Process Control)

Statistical process control (SPC)
Statistical process monitoring (SPM)
Multivariate SPC or SPM
Real-time SPC
Equipment monitoring
Excursion detection and control
Fault detection (or fault identification)
Fault isolation
Fault classification
Diagnosis
Fault prognosis

TABLE 23.2 Terms and Groupings of Process Control Methods That Are Based upon Compensation for Expected Variations (Automatic Process Control)

Model-based process control
Sensor-based process control
Engineering process control
Algorithmic SPC
Automatic process control
Automated process control
Feedback/Feedforward control
Specific types of control algorithms fuzzy logic, model predictive control, robust control, etc.
Run-to-run control (batch to batch)
Real-time control (within a batch)

compensation control methods were first developed, the compensation was determined by running calculations on computers. Thus, these techniques can also be considered automatic control. This idea of compensating for expected variation can be seen in Figure 23.2, where the goal of the controller is to reduce the variance around the target to the inherent random noise in the system. A concept that Box and Hunter have been introducing in recent presentations is that compensation control transfers the variance from the output, where it is expensive, to the input, where it can be less expensive [3]. Thus, compensation control results in cost savings.

The groups of methods listed in Table 23.2 vary due to the source of data used by the method, the types of algorithms, and the time scale over which the algorithms run. Within a group, the methods vary due to the use of different algorithms for deciding the compensation. Different groups of methods are more appropriate for some types of variations. The compensation control methods all decide (1) when to make an adjustment, (2) what variables to adjust, and (3) how much to adjust those variables so that the desired results will be achieved. Different methods do a better job of compensation based upon the particular source of variation, the dynamics of that source, the random noise level, and the particular controller goals that the engineer desires. Tracking the target is typically the predominant controller goal, and thus the term "target tracking" to refer to compensation control methods. However, other goals exist in addition to target tracking [4].

23.2.3 Advanced Process Control: Combination of Both

Compensation control methods should also be aware of unexpected variation (or abnormality). This situation is when the controller operates on a system with different variances than for which the controller was designed. In other words, the random and/or non-random variances are different than expected. Although the controller may still drive the output to target, the overall quality of the result is suspect since operating in this regime has never been qualified. Thus, unmeasured variables and the metrology itself may be out of specification. The desired controller behavior would be for the controller to detect the change in process behavior and generate an alarm. Note that this situation can also arise in the abnormality based methods when the output appears to be on target but in reality the system has drifted to a state where the metrology and the system are out of specification. In this case, abnormality-based methods have no extra ability to detect the system change. Fortuitously, the compensation behavior provides another way to determine if the system is behaving as expected. In other words, the compensation based controller can monitor not only the output changes, but how the output changes

TABLE 23.3 Alternative Terms for Run to Run Control

Run-to-run (run-to-run) control	Run-by-run (run-by-run) control
Supervisory control	Batch-to-batch control
Recipe adjustment	Recipe generation
Recipe synthesis	Recipe tweaking

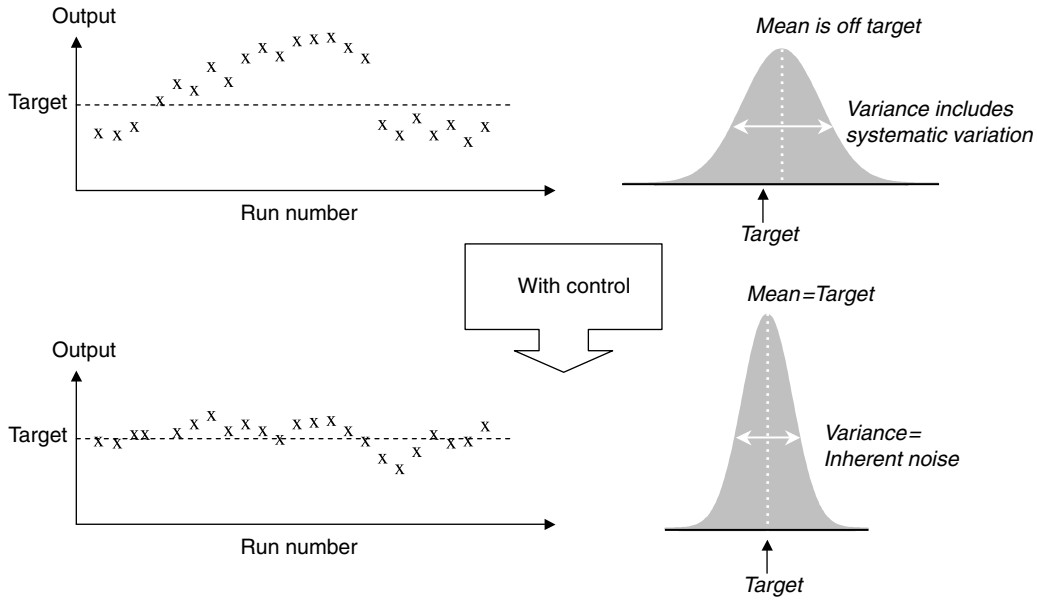


FIGURE 23.2 Driving the output mean to the target and shrinking the variance to the inherent random noise.

in response to a given change in input as well, to determine if the over-all system has changed behavior resulting in better detectability. This desire to monitor for abnormalities in the compensation based controller leads to the concept of merging both methods into one.

The merging of both compensation and abnormality based methods is called Advanced Proess Control (APC) in the semiconductor industry. The supposition in APC is that there is expected non-random (systematic) variation, expected random noise, and unexpected variation (which may be non-random or random). Compensation is assumed desirable and possible for the expected non-random (systematic) variation. Unexpected variation is to be detected and an alarm generated preventing the controller from operating under these conditions. Thus, a "fault" is a change in the variation of the system away from the expected variation. In other words, APC decides:

- When to make an adjustment
- What variables to adjust
- How much to adjust those variables.
- If the system is responding and acting as expected, such as by checking that adjustments are not to frequent or too large or that the system has not drifted too far.

Because of the number of decisions made in APC, it is also defined as a documented analysis methodology for deciding how to run a single piece of equipment or a group of equipment to achieve desired results. Another view of APC is shown in Figure 23.3 in which the various components which run at different time scales are shown. The various controllers, monitoring systems, and data sources are shown. Manufacturing enterprise system (MES) which stores the product and process specifications, as well as tracks work in process (WIP). The Recipe is the setpoints and parameters for all the controllers on the equipment, i.e., it tells the equipment how to operate. Each layer builds upon the previous layer by using data from the previous layer plus additional information to achieve that layer's goal of abnormality detection or compensation. Metric generation is explicitly shown in Figure 23.3 due to its importance. To emphasize that it is the controller data that is monitored to detect a system fault, the abnormality detection part of APC has also been termed "statistical process control (SPC) on the controller" or "controller SPC."

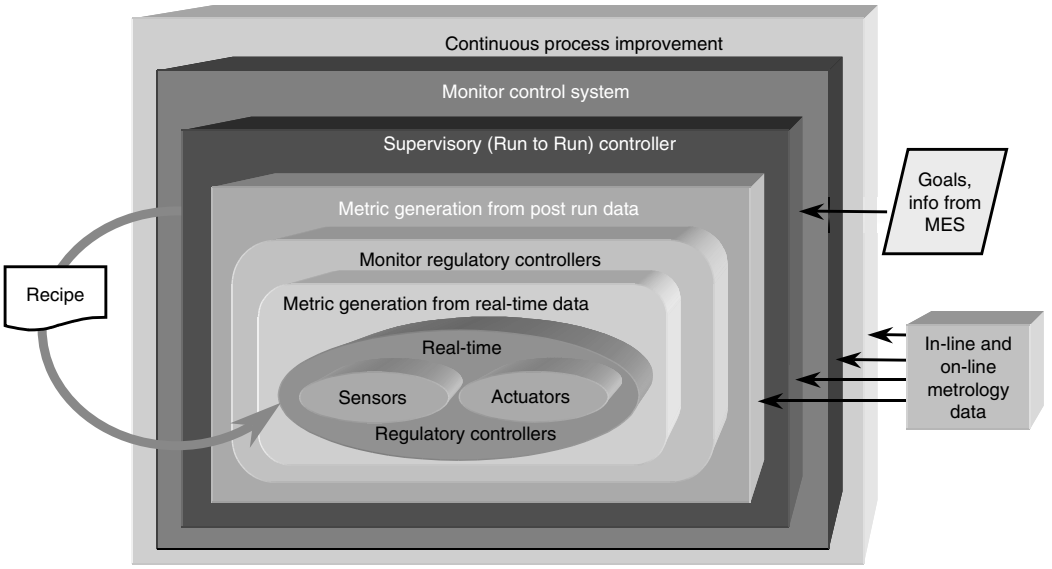


FIGURE 23.3 Components of advanced process control (APC): Optimum performance by integration of each layer. Real-time regulatory control includes endpointing; MES is manufacturing enterprise system (also known as computer integrated manufacturing system); supervisory run-to-run controller has historically been referred to as "model-based process control"; monitor for regulatory controller has historically been referred to as "fault detection and classification"; monitor for supervisory controller has historically been referred to as "statistical process control on the controller" or "overseen."

It is also common to see APC define as:

$$\text{APC} = \text{MBPC} + \text{FDC}$$

where MBPC is model-based process control, which is commonly used for the supervisory run-to-run controller and FDC is Fault Detection and Classification, which is a common generic term for monitoring the regulatory controller. It also implies Fault Classification and Fault Prognosis.

A similar concept, algorithmic Statistical Process Control, was being investigated at approximately the same time in the process industries [5,6]. The major difference is that Algorithmic Statistical Process Control focuses only on controlling the "quality variables," i.e., only the outer three layers with ex situ data of Figure 23.3, i.e., the layers utilizing at-line and on-line data for supervisory control and monitoring of the supervisory controller. A similar philosophy of merging the best of SPC and automatic control was also being investigated by a few in the statistical and control communities [7–14]. In these investigations, it varied whether the focus was on real-time or run-to-run control. However, generally, the focus was on only one of the loops (real time or run-to-run), rather than synergistic linkage of compensation control and abnormality detection at all levels.

23.3 History of Process Control in Semiconductor Manufacturing

While real-time compensation control techniques had existed for a few decades in the petrochemical, process, and aerospace industries, the development of control systems in the semiconducting industry initially was not based upon work in these other industries. While some of the academic and industrial control people were involved in more than just the semiconductor industry, an explanation of the history

will show that the independence was due predominantly to the systems environment lack of real-time process and product sensors on commercial equipment, focus on Statistical Process Control, which evolved into run-to-run (batch-to-batch) control, and real-time controller monitoring. Initially most semiconductor processing equipment had very little process control other than temperature controllers on furnaces. However, by the mid to late 1980s, pressure controllers, mass flow controllers, and temperature controllers were becoming quite common on most tools. Lithography exposure equipment had better controllers for dose, focus, and alignment. Feedback control of unit processes was now being recognized as necessary for consistently achieving future smaller device geometries.

As part of this recognition, the Semiconductor Research Corporation (SRC) and SEMATECH began focusing funding in the area of process control. Control was considered part of Computer Integrated Manufacturing (CIM) or Factory Science, and therefore this topic was included in the CIM and Factory Sciences Areas. An SRC Workshop (first CIM workshop) was held at the University of California at Berkeley in 1986. A second workshop was held the next year at the Massachusetts Institute of Technology. These workshops helped focus attention on process control. An SRC workshop on Real-time Tool Control was held on February 1991 in Canada, co-hosted by Techware (now Brooks Automation Canada). A SRC/DARPA CIM workshop was held in August 1991. SEMATECH then began hosting Advanced Equipment and Process Control workshops from the summer of 1991. The list of workshops is included in Table 23.4.

Initially, these workshops focused on real-time process control. However, due to the lack of commercial real-time process sensors, the apparent huge hurdle to be overcome by process equipment suppliers, and a very strong interest by semiconductor manufacturers, the focus of the workshops became more increasingly focused on run-to-run control and regulatory controller monitoring. Semiconductor manufacturers could implement these controls without involvement of the equipment supplier. In addition, large benefits could be achieved in both these areas. While the workshops were driving the academic, industrial, and equipment research communities, a revolution was occurring on the factory floor. The need for increased control and the incompatibility of traditional Statistical Process Control (SPC) with many semiconductor processes was resulting in the invention of compensation control and advanced abnormality detection methods by fab process engineers [15]. So, although the typical fab process engineer was not versed in control theory, they also were working in areas traditional control theory mainly did not exist: run-to-run (batch-to-batch) control and regulatory control monitoring (batch monitoring). In addition, statisticians heavily assisted the fab engineers, a situation unusual in the process and petrochemical industries. The result was a different theoretical background for compensation control. Finally, the strong link to the Manufacturing Execution System (MES) was not present in other industries. Thus, the unique systems environment hindered the application of traditional process industries control theory to the semiconductor industry, while simultaneously enabling the development of this new area of run-to-run control. In addition, because of the strong Statistical Quality Control fab environment, the integration of various levels of control resulted in the concept of Advanced Process Control (APC) presented in the last section.

TABLE 23.4 SEMATECH Advanced Equipment and Process Control Workshops

I	30 July to 1 August 1991, Austin, Texas
II	3–5 March 1992, Mesa, Arizona
III	October 1992, Austin, Texas
IV	19–22 April 1993, Dallas, Texas
V	October 1993, Dallas, Texas
VI	Fall 1994, San Antonio, Texas
VII	Fall 1995, New Orleans, Louisiana
VIII	Fall 1996, Santa Fe, New Mexico
IX	Fall 1997, Lake Tahoe, Nevada
X	Fall 1998, Vail, Colorado

Title of workshop has changed through the years.

By the mid-1990s, traditional and modern control theory relevant for run-to-run control was being integrated into the fab-invented and statistically-based techniques [16,17]. In addition, the regulatory control monitoring (equipment signal monitoring) theories from semiconductor manufacturing and process industries were being cross-fertilized. Also, changes at process equipment vendors and the emergence of commercially available process sensors were resulting in the re-emergence of real-time control focus. While the control theory from other industries is being exchanged, the unique systems environment has resulted in unique systems solutions for the semiconductor industry. For further explanation of the systems issues, see Ref. [18].

Because of the emphasis in the industry on run-to-run control, regulatory control monitoring, and the concept of Advanced Process Control (APC), these areas will be more heavily focused upon in this chapter. The ability to understand exactly what are the control needs, i.e., why are abnormality and compensation control methods needed, is imperative to deciding which control technique is appropriate. This understanding will also provide insight into how control evolved in the industry since it was focused on meeting the needs of the semiconductor industry. The next section will present a dissection of the different control needs in the semiconductor industry. This dissection is based upon the historical observation of what are common control needs process to process and fab to fab.

23.4 Characterization of Control Needs in Semiconductor Manufacturing

The benefits of control have been observed by many and presented in a variety of articles, as well as at the SEMATECH workshops [19,20]. Table 23.5 provides a list of benefits that can be achieved with the right control method. The benefits can be considered direct and indirect. Direct benefits impact the cost or quality of manufacturing in an easily observed manner. Indirect benefits lead to cost or quality improvements in manufacturing, sales, or design in a less than obvious manner. There are other benefits which are not listed in Table 23.5 because they are software dependent. Two such software-enabling benefits are the removal of paper logbooks from the fab, and the integration of qual data and producing data into our database.

All the benefits listed in Table 23.5 can be lumped into three categories as shown in Figure 23.2 and Figure 23.4. The compensation controllers continuously drive the output to target, thereby improving

TABLE 23.5 Detailed Benefits due to Advanced Process Control

Direct Benefits	Indirect Benefits
Reduced pilot and look-ahead usage	Improved productivity of operators and engineers
Reduced number of quals and qual time	Improved customer satisfaction
Reduced set-up time	Continuous improvement of equipment (Pareto sources of variation)
Reduced cycle time (due to qual and set-up time)	Faster learning cycle for process development (faster device ramp)
Reduced scrap (outlier reduction)	Process and product understanding
Accommodate inherent machine/chamber difference	Accommodate greater device diversity (increased process flexibility)
Reduced (tighter) distribution around target	Fab-to-fab standardization and sharing
Increased equipment uptime and productivity	Increased confidence in alarms
Decreased equipment repair time	Increased ability to redesign products based on tighter process control
Decreased post-maintenance recovery time	
Improved data-driven maintenance and qual schedules	
Improved qualification of equipment kits	
Reduced capital costs (increased life of equipment)	
Data drive sampling plan	

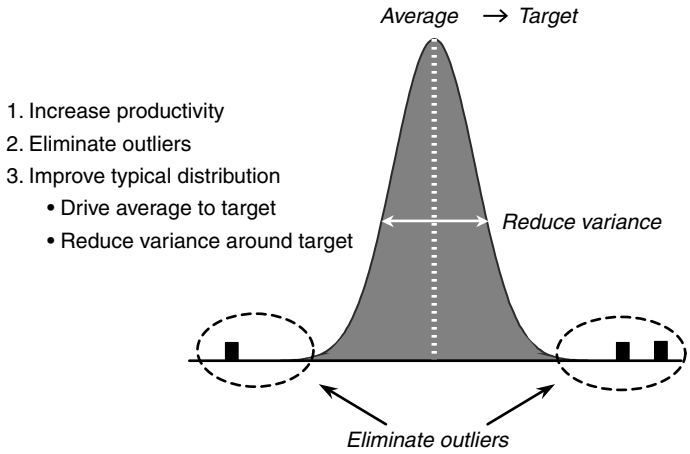


FIGURE 23.4 Outliers vs. expected typical distribution and ways in which APC results in improvements.

the distribution. The abnormality based methods detect outliers and prevent continued misprocessing. When both methods are used, as in Advanced Process Control, the sensitivity to changes is increased, and the number of outliers is reduced further. The productivity improvement is fortuitous and is due to the automation, applicability of the mathematics, and heavy emphasis on decision-making of APC. As discussed [19,20], it is the productivity improvement that is most needed. We will discuss these benefits again in the Section 23.7.1.

As noted in the introduction, different methods perform better for different situations, and applying appropriate statistics is important to achieve the needed benefits [21]. Consequently, a good understanding of the situation in which the methods will be applied is required to understand which method will work satisfactorily and why the benefits listed in Table 23.5 and Figure 23.4 can be achieved. Also some of the productivity benefits are due to the way pilots (non-sellable wafers) and metrology are used in the semiconductor industry.

In this section, the reader should consider whether the variation is:

- Expected Random Variation
- Expected Non-Random Variation
- Non-expected Variation

As mentioned in Section 23.2, compensation methods are aimed at correcting for expected non-random variation and abnormality-based methods are aimed at detecting non-expected variation. For both types of methods, expected random variation is considered natural and should not be classified as a fault, nor should compensation be attempted.

23.4.1 The Four Expected Sources of Variation Requiring Compensation

The variation for which control compensates can be categorized into four sources. Table 23.6 Provides the four categories of expected variation for which control is compensating. The sources of variation in Table 23.6 require compensation, i.e., a different action must be taken, in order for the desired results to be achieved. Note what is not included in Table 23.6. An unstable process is not included. A machine that breaks frequently is not included. Thus, while compensation is a tool that every engineer should have in their toolkit, it is not the only tool.

Dynamics can be defined as the non-random behavior of the system over time, i.e., how the output would change with each run if no compensation were used. Disturbances themselves have a particular

TABLE 23.6 Four Expected Sources of Variation Requiring Compensation with Explanation and Examples

1. Within-run dynamic process or dynamic target
   → Strong function of time within a single run
2. Disturbances[a]
   → Equipment aging, such as calibrations, chamber build-up, part wear
   → Machine maintenance, such as chamber clean, kit replacement
   → Inherent chamber-to-chamber differences
   → Wafer state (called "loading" in etch)
3. Change in feedforward values[b]
   → Incoming wafer state, which are due to results of earlier processes, such as thickness
   → Machine attributes, such as tube age, sputter target age
4. Change in process or product goal (e.g., set point, target)
   → Different device
   → Different step in flow (sidewall oxide etch as opposed to contact oxide etch)

Usually not want to compensate for ALL disturbances, feedforward changes, just expected ones.

[a] A disturbance which can be measured and whose impact on the controlled output can be modeled can become a feedforward variable.

[b] An unmodeled or unmeasured change in a feedforward variable is a disturbance.

dynamic behavior which are given in Table 23.7. Feedforwards and Goal changes can be viewed as step changes, i.e., dynamics so fast as to be unmeasurable. Understanding the dynamic behavior is important because the compensation technique is dependent upon the specific dynamics of the system. In other words, what is called the systematic variation by the process engineer is called the dynamic behavior by the control engineer. A technique for step dynamics may not work as well for a disturbance whose dynamics are a moderate ramp. The size of the change is also considered when designing the compensation technique. If the change expected is large, more aggressive action may be required than if small changes are encountered.

23.4.2 Time Scale of Variations

As the same theme as dynamics, is the time constant of the system. In other words, the amount of time it takes before changes are observed. There are actually several different sources of variation within

TABLE 23.7 Equivalent Dynamics (→) and Size of Change (❖) for Disturbances

1. Equipment aging, such as calibrations, chamber build-up, part wear
   → Slow-to-moderate run-to-run dynamics
      ❖ The faster the dynamics, the greater the changes over 10 lots
2. Machine maintenance, such as chamber clean, kit replacement
   → Step function
      ❖ Size may range from small to very large, typically large
3. Inherent chamber-to-chamber differences
   → No dynamics
      ❖ Size may be significant, but can be driven to near zero with considerable work
4. Wafer state (called "loading" in etch)
   → Step function as switch between products due to differences in topography, open area
      ❖ If exists, size can be considerable
   → May be step function or slow dynamics if due to previous equipment disturbances
      ❖ If exists, effect is smaller than the effect due to different products
5. Major fault, such as mass flow controller (MFC) fails or rapidly degrades
   → Step function or within-run dynamics (fast run-to-run dynamics)
      ❖ Size may range from small to very large, typically large

A disturbance that can be measured and whose impact on the controlled output can be modeled can become a feedforward variable. Even if the disturbance becomes a feedforward, the dynamics still exist.

a semiconductor process system. Each of these variations and their associated time scale are given in Table 23.8. Table 23.8 assumes a single wafer processor, i.e., that each run consists of a single wafer. Not all systems are single wafer, such as batch furnaces. However, because single wafer systems represent the most complexity, only they will be analyzed. Note that there are continuous changes and discontinuous changes. For example, between each maintenance, there is the continuous slow drift of the order of a fraction of the maintenance cycle, i.e., the changes can only be observed over 100 or more wafers (actually over several lots). However, there are also discontinuous changes which occur approximately every lot (24 wafers). The same continuous and discontinuous variations occur when observing what occurs within a time scale of 24 wafers. Finally, changes occur within a single run itself. Figure 23.5 presents a plot of an example of within run variation. This example is metal etching in a plasma. While there is considerable variation within a single run, the functionality is repeatable. In other words, the mean value or length of a region may increase or decrease from run to run, but that one region has a higher mean value than another region will repeat from run to run. The length of the region may vary due to different incoming or targeted film thicknesses, as well as process drifts over larger time scales (e.g., Within a Maintenance Cycle). The mean value may change due to sensor aging, process aging, or changes in incoming wafer state. The SMALLEST size of variation is typically the within a lot, wafer to wafer, and within a region.

TABLE 23.8 Dynamics of Each Time Scale Variation and Their Causes (Assumes a Single Wafer Processor)

- Maintenance cycle-to-maintenance cycle (every $>$20,000 wafers)
  - Discontinuous change (step function)
  - Repairs, chamber cleans, preventive maintenance kit replacements
  - Attempts to return in same ideal machine state are rarely successful
  - *Maintenance can produce the largest changes*
- Within a maintenance cycle ($\sim$100–1000 wafers)
  - Continuous change
  - Gradual build-up on chamber, machine ear, sensor drift
- Lot to lot (every 24 wafers)
  - Discontinuous change (step function)
  - Due to incoming wafer state
  - Due to current process (if plot data from lots of a given type)
    - Other lots processes run between lots of this process
  - *Undocumented* maintenance, *changes* generally occur between lots
- Within a lot [assuming one chamber (1–10 wafers)]
  - Continuous change
  - "First wafer effect" (which may last for more than one wafer)
    - Warm-up effect, de-gassing, different steady-state chamber state
  - Due to incoming wafer state from previous process which has first wafer effect
- Wafer to wafer (every one wafer)
  - Discontinuous change
  - Due to incoming wafer state
    - For example, left/right track effect of upstream lithography step
    - Different chambers of cluster tool used for upstream processing (1–2–3–1–2–3…)
  - Due to current process
    - Randomness caused by process start-up, repeatability of equipment controllers
  - *Undocumented* maintenance, *changes can also occur wafer to wafer*
- Within a wafer, further broken down into (Figure 23.5)
  - Region-to-region
    - Discontinuous change
    - Due to different processing conditions (different recipe step) or due to different materials exposed
  - Within a region
    - Continuous change
    - Due to changes in materials exposed and heating effects

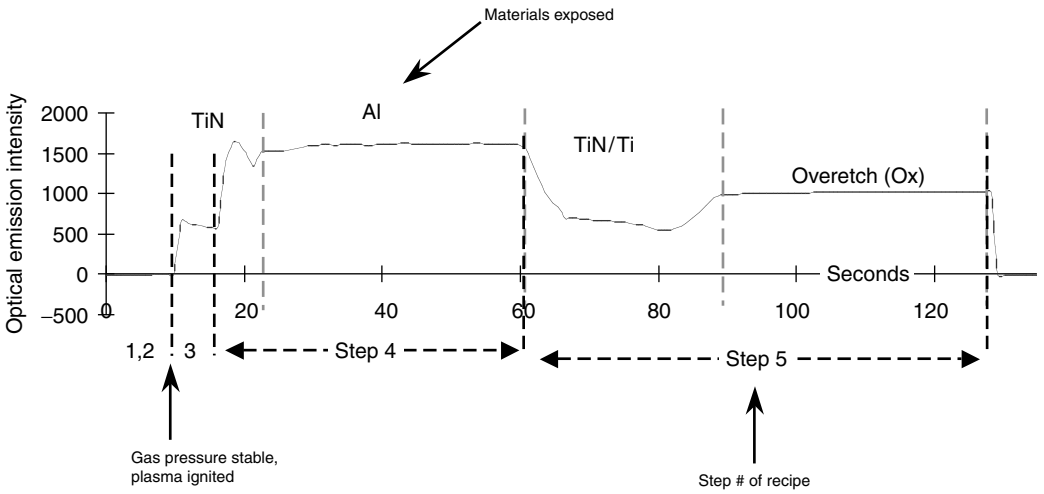


FIGURE 23.5 Example of within a wafer variation: metal etch.

It is imperative to understand the main sources of variation for both fault detection and control. Many years of observation have shown faults generally occur between lots. Faults are next most likely to occur between wafers. One wants to catch the fault on or before the first wafer after the fault occurred. However, this can be very challenging considering large variations are expected to occur between lots, and the fault may appear small to the measurement technique. Thus, detection between wafers is easiest, but can ONLY occur if measurements are made on every wafer. In-line metrology rarely is done on every wafer. Consequently, only in situ or on-line metrology (metrology mounted on the process equipment or equipment sensors), which can easily be used for every wafer, can be used to detect faults between wafers. An additional challenge is caused by the variety of processes run on a single tool. The difference between processes is usually much larger than the size of a fault. Thus, one might propose to analyze each process separately to increase sensitivity to the fault. However, due to all the expected variations which occurs to the machine and that a given process will not be run in huge volume, the variability across lots of a single process can be huge too. Therefore the concept of global modeling, where a model is used to normalize all processes to each other, is used to allow increased sensitivity. This concept is also used in feedback/feedforward control and is the predominant source of productivity improvement due to elimination of pilots.

Understanding the time scales is important for control because they are a part of the dynamics analysis. As discussed above when introducing dynamics, the compensation technique is selected based upon the dynamics. Thus, if the major source of disturbance is machine variation caused by maintenance a different control scheme will be designed than if the major source of disturbance is the first wafer effect.

23.4.3 Pilots, Look-Aheads, Metrology, and Operational Practices

Twenty to Forty percent of wafers processed are pilots (non-production wafers). Another significant fraction are look-aheads, wafers from a production lot processed and analyzed before the rest of the production lot is processed. Pilots cannot be sold and look-aheads have a huge impact on cycle time and throughput. Thus, both are undesirable. One reason for pilots is that some of the current wafer metrology is contaminating, destructive, or cannot work with topography. Another usage of pilots is for chamber conditioning post maintenance. However, manufacturing flexibility and lack of direct measurement of the equipment/chamber state are the actual reasons for the bulk of pilot and look-ahead

usage. Manufacturing flexibility mandates that a single piece of processing equipment process a variety of wafers. Thus, concern arises about how one process (called "A") will impact the chamber which results in a different result for another process (called "B"). Without any on-tool measurements which indicate the chamber state, a wafer measurement must be used to indicate the chamber state. Therefore, when switching between processes A and B, a pilot or look-ahead will be run first to confirm the chamber state. If there is confidence that process B will most likely be ok or only need small tweaking, a look-ahead is used. If there is less confidence, a pilot is used. The same dilemma arises after maintenance, will the chamber be different and cause different process results? Again, pilots or look-aheads are used. Regardless of process interactions, because processes A and B may be very different, without advanced control, processes A data was analyzed separate from process B data. Thus, if the last 10 lots run were all of process A, a pilot or look-ahead is run before process B. Not because A is suspected to have cross-talk with process B, but because no data are available to predict how process B is run. If a global process model can be created to relate process A data to process B data, then a look-ahead need not be run. It is this ability to create a global model that reduces the number of pilots used in a production fab. In some instances, if process A and B interact, a global model may be developed which predicts this interaction. Modeling and understanding the interaction is much more difficult. Thus, interacting processes are more infrequent in the fab, while non-interacting processes are quite common. While advanced control can reduce the pilot usage for machines which run more than one process, only better on-the-tool metrology can reduce pilot usage after maintenance.

23.5 Basic Concepts of All Control Techniques

Before discussing specific techniques, there are several concepts that the reader must comprehend. The first is the Process Qualification and how it pertains to control. The next concept is the Process Capability Indices and how they are used to rank control. The final concept is error rates and their usage in tuning the control algorithm.

23.5.1 Process Qualification

Because this chapter is on process control, to not discuss the role of process qualification would be negligent. Thus, before specific control techniques are presented, a methodology for identifying the right control technique and insuring its ability to control the process satisfactory will be introduced. Included is an over-all semiconductor processing quality system is a methodology for qualifying a process for use in manufacture of production material. The methodology for qualifying a process is commonly known as a qual plan. The most well documented Qual plan in the open literature is the SEMATECH Qual Plan [22,23] Domain Solutions markets software to assist in conducting a qual plan, based on one created at IBM [24]. An important part of a Qual plan is to determine that the associated metrology is capable. Gauge studies, also known as R&R studies, which signify the analysis of repeatability and reproducibility variances, are used to assess the metrology capability. Variance component studies are also done of the process to determine that it will meet specifications with the desired yield and determine the source of variability. This study is done with the desired control system in place. Thus, evaluating the quality of the controller is part of qualifying a process. An adequate control system will guarantee that methods and practices exist which insure the capability values will be maintained in the actual production environment. In the development of the controller, the sources of variation that will be encountered must be considered to design an adequate controller.

23.5.2 Process Capability Indices

Process capability indices are used to assess the process' ability to achieve yield. The two most common over-all metrics to assess the process's capability are C_p and C_{pk}. These indices were created in the statistical Quality Control field. However, they are a useful metric even when evaluating compensation controllers.

C_p strictly evaluates the process's variability compared to the specification limits on that process:

$$C_p = \frac{\text{USL} - \text{LSL}}{6\sigma} \tag{23.1}$$

where USL is the upper specification limit, LSL the lower specification limit, and σ the standard deviation of the process.

While C_{pk} also considers the mean of the process, i.e., how centered the process is within its specification limits:

$$C_{pk} = \text{minimum}(C_{pL}, C_{pU}) \tag{23.2}$$

where

$$C_{pU} = \frac{|\overline{X} - \text{USL}|}{6\sigma}$$
$$C_{pL} = \frac{|\overline{X} - \text{LSL}|}{6\sigma}$$
$$\overline{X} = \text{Average of process results}$$

The desired values of C_p and C_{pk} are commonly said to be 2 for a 6σ process. Figure 23.6 shows a graphical representation of C_p and C_{pk} with a value of 2, along with a shift in the mean of 1.5σ yielding a C_{pk} value of 1.5. Note that when the mean shifts and the variance does not, as shown in Figure 23.6, the C_p value remains unchanged, as also shown in the Figure 23.6. A C_{pk} of 2 means that only 2 parts per billion (ppb) are outside of the specification limits, while a value of 1.5 means that 3.4 parts per million (ppm) will be out of specification. Note that 2 ppb loss translates to a yield of 99.9999998% and 3.4 ppm loss equals an yield of 99.99966%.

Because of the many statistical problems exist with C_p and C_{pk}, such as assuming both an upper and lower specification limit and only one source of random Gaussian variability, there is considerable work in the statistics field developing alternative process capability indices. The greatest issue is the lack of understanding of the large sample size required to obtain a value with good confidence, but this situation

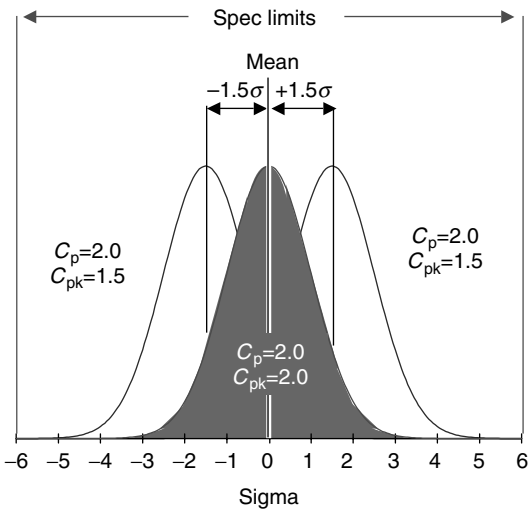


FIGURE 23.6 C_p and C_{pk} process capability indices; C_{pk} considers the average value while C_p does not; C_{pk} value of 2 is equal to 2 ppb outside of specification; C_{pk} value of 1.5 is equal to 3.4 ppm out of specification.

would be true for any metric involving variance. Regardless of the statistical issues, C_p and C_{pk} provide engineers a target for which to aim, and they also convey the importance of centering the process and shrinking its variability.

Two alternative metrics for judging a process's performance were developed in the process industries control area, which are independent of the specification limits [4].

$$\text{MISE} = \sqrt{\frac{\sum_{i=0}^{t} \text{error}_i^2}{t}} \tag{23.3}$$

$$\text{MIAE} = \sqrt{\frac{\sum_{i=0}^{t} |\text{error}_i|}{t}} \tag{23.4}$$

where error_i is the target-Y at time i, Y the controlled output, MISE the mean integral squared error, and MIAE the mean integrated absolute error.

These two metrics are influenced by how centered a process is and the variability of the process. MISE is the analog of the standard deviation, but with the average replaced with the target. In other words, MISE can be considered the standard deviation around the target. MIAE is the analog of the Mean Absolute Deviation, and just like Mean Absolute Deviation, is less sensitive to outliers. Thus, process modifications improve MISE greater fraction than MIAE if the modifications remove outliers ("spikes"). These two metrics do not convey information on yield, but they do indicate impact of process and equipment modifications. The goal is to continuously shrink the values of MIAE and MISE.

23.5.3 Types of Errors: False Positives and False Negatives

The following discussion is based upon linear statistics for white gaussian random variation, but these simplified concepts are applicable to any situation. Let's assume that one is attempting to detect a shift in the mean only with no change occurring in the variation level around the mean. As will be discussed later, this situation is the one usually assumed in abnormality control methods. A sample of size N will be used to detect the shift. Consequently, there are two probability distributions of concern, the original distribution and the distribution shifted by an amount Δ. The two distributions are shown in Figure 23.7. A t-test is used to determine if a shift has occurred [25]. Table 23.9 defines the two types of errors which will occur with the costs associated with each of the errors. As Figure 23.7 shows, the overlap of the two distributions results in the two types of error. It can be seen that as Δ decreases, β will increase if α is held

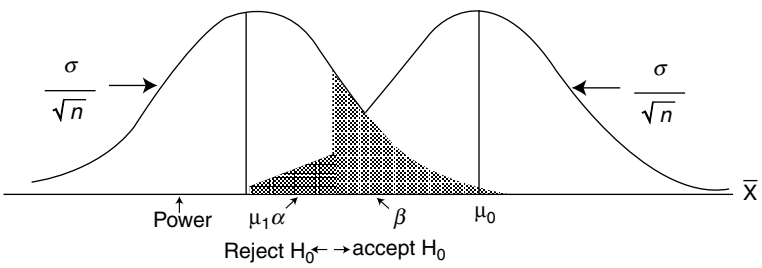


FIGURE 23.7 Type I (α) and type II (β) errors from the comparison of mean values of two normal probability distribution functions for sample of size N. Mean values are a distance of Δ apart. Standard deviation (σ) is the same for both distributions.

TABLE 23.9 Types of Errors

Type	% of Errors of that Type	Indicates	Description	Costs Due to
I	Alpha (α)	False positive	Detect an abnormality when it has not occurred	Wasted time trying to track down fault that does not exist; unnecessary machine downtime
II	Beta (β)	False negative	Not detect an abnormality when it has occurred	Shipment of bad product, lost of quality, loss of improvement opportunity due to inability to see impact of a factor

constant. In other words, for a given sample size, smaller shifts are more difficult to detect. As the sample size N is increased, the distributions narrow since the standard deviation of the sample mean is the standard deviation of the individuals divided by the square root of N. Thus, as the sample size increases, for a constant shift of size Δ, β will decrease if α is held constant (or vice versa). In other words, increasing the sample size makes a shift easier to detect.

To summarize, Table 23.10 lists the variables which determine the percentage of type I and type II errors, α and β, respectively, for any fault detection method. Which mathematical technique is selected is based upon its appropriateness to the situation and the availability of software and computer hardware. The size of fault that needs to be detected is based upon the necessary quality that must be achieved. However, for the mathematics, it is not the absolute size of the fault that matters, but rather the relative size of the fault compared to the normal variation, usually represented by the standard deviation. The greater the sample size, the lower are the type I and type II errors. However, larger sample sizes may take longer time to gather which may increase the amount of time before a change will be detected. In addition, many times, the sample size is determined by the physics and cost of sampling. Consequently, the remaining choice is between α and β. Once a value for ONE of these variables is selected, the value for the outer variable is also determined. As described in Table 23.9, there are costs associated with both of these errors. Thus, the trade-off between α and β is really an optimization of the economics.

The value selection for the error types, for a given relative fault size to detect and sample size, is determined by how the mathematical technique itself is set-up or tuned. In other words, each fault detection technique generally has at least one parameter associated with it. A given value for the parameter(s) determines the values for α and β for a given sample, size, normal variation level, and desired fault size to detect. Many of the methods also include sample size as an explicit parameter in the equations, further highlighting the importance of the sample size.

Another way error type information is calculated and explained is with the average run length (ARL). This metric is useful for cases where analytical calculation of error rates is difficult. The ARL is calculated by performing many simulations for a given fault size and standard deviation of the noise ("common cause variation"). The number of runs before a fault is detected, the run length, is recorded for each simulation. The average of these run lengths is then calculated. The ARL for a fault size of zero provides a metric for type I error. The ARL for a given non-zero fault size provides a metric for type II error. Thus, a chart or table of ARL vs. fault size per normal standard deviation is used to compare different fault

TABLE 23.10 Variables Involved in Determining α and β Error Rates

- Mathematical technique used
- Size of fault being detected when compared with normal variation
- Type I error (α)
- Type II error (β)
- Sample size of the data

detection methods and/or different values of the method's parameters. While rarely done, the whole distribution of run lengths could also be compared.

23.6 Specific Abnormality Detection and Control Methods

We will cover abnormality detection methods used in run-to-run control under compensation methods. Generally, abnormality detection and control methods are part of a total quality management (TQM) program [26]. Due the volume of existing textbook on SPC and TQM, TQM will not be covered here and SPC methods will only be quickly reviewed.

23.6.1 Univariate Statistical Process Control

The most common of the abnormality-detection-based methods in the semiconductor industry is Statistical process control (SPC). There are several books dedicated to the application of SPC [27]. SPC has been practiced in the semiconductor industry for at least 20 years and is very widespread [28]. The SPC is an entire methodology, including addressing what actions to take upon detection of an abnormality and how to inform the operators of these needed actions.

In traditional SPC, the expected variation is assumed to be described by a normal (Gaussian) distribution occurring around a mean. In other words, the errors around the mean are assumed to be identically independent distributed (IID) normal. Identically independent distributed means that the value of each error for every measurement comes from the same distribution, and that each value is independent of the previous value, i.e., the errors are uncorrelated. This assumption for the distribution of y is represented as

$$y = \mu + \varepsilon \tag{23.5}$$

$$\varepsilon = \text{IIDN}(0,\sigma) \tag{23.6}$$

where y is the measurement, μ the mean (average) of the distribution for y, ε the random error in measurement y, IID$N(0,\sigma)$ the Identically Independently Distributed Normal (Gaussian) distribution with mean of 0 and standard deviation of σ.

Another common way of representing the distribution of y is

$$y = N(\mu,\sigma) \tag{23.7}$$

where y is the measurement, $N(\mu,\sigma)$ the normal (Gaussian) distribution with mean of μ and standard deviation of σ.

In SPC, an abnormality is assumed to be a shift in the mean of this distribution (μ) or a change in the standard variation of the normal distribution (σ). In SPC, the abnormality detection technique used is based on the statistics and charting of the data. Different types of statistics have a different associated charting method. Thus, the specific fault detection techniques are usually called XYZ chart, with XYZ denoting the specific statistics used. Note that many times there may be more than one actual chart per a given technique. Different techniques exist because different physical situations require different mathematics and due to efforts to obtain better type I and type II errors. Also, different techniques may be testing different hypothesis. Some are testing whether the mean (μ) has shifted, while others are testing whether the standard deviation (σ) has changed. Due to the statistics, it requires a much larger sample size to detect a change in the standard deviation than a change in the mean [25]. In addition, some may argue that changes in the mean are more likely to occur, although we will address this suggestion again in the section on using equipment signals. Consequently, charts to detect changes in the mean are much more common. Incidentally, a change in the variance can also cause a test for a change in

the mean to trigger. Thus, although the fault detected is incorrect (a mean change when it is a variance change), an alarm is still generated indicating that something is abnormal.

The most common of the SPC charts is a Shewhart chart, also known as an XBar–R (Average–Range) chart [29]. Although XBar–R charts are the must well known, the I-MR (Individuals-Moving Range) is more appropriate. The across sample and run to run variation must be the same for an XBar–R Chart to be appropriate. The variation within a wafer, wafer to wafer, or lot to lot is rarely the same, thus causing the use of a single sample ("individual"). To decrease type II error, "supplementary run rules" are used with the Shewhart Chart. These rules are generally known as the Western Electric (WECO) Rules in recognition of the source of their well known application. While WECO rules decrease Type II errors, they also increase Type I error. Other charts, such as CUSUM, have been developed to provide better Type I and II errors for smaller relative fault sizes. These other charts also lend themselves naturally to single sample size applications. However, these alternative chart types still have not achieved wide application mainly due to software limitations.

23.6.2 Fault Detection by Testing That There Is a Fault

Note that traditional Statistical Process Control Methods look for an abnormality by testing the hypothesis (H_0) that the system IS as expected. For example, to test that the current mean equals the expected mean:

$$H_0{:}\mu = \mu_0 \tag{23.8}$$

where μ is the current mean and μ_0 the expected mean.

On the other hand, one could explicitly test that the system is NOT as expected by testing for a particular fault, for example:

$$H_0{:}\mu = \mu_0 + \Delta \tag{23.9}$$

where Δ is the fault size testing.

For Equation 23.8, only data from when the system is "good" are required to set-up the fault detection method. However, to actually understand the α and β values, a fault size must be assumed. For Equation 23.9, data from when the system is "good" and from when the system is "bad," i.e., experiencing the fault, are required. Thus, if a given fault is known to occur, the methodology corresponding to Equation 23.9 will generally yield better fault detection capabilities. However, if no particular fault is expected, the methodology represented by Equation 23.8 is the only approach available. Even if Equation 23.9 is used, Equation 23.8 should also be used to catch all unexpected faults. While this concept of using "bad" data to create an abnormality detection method is beginning to be presented more frequently in the literature, it still is not encountered in industrial practice.

23.6.3 Other Abnormality Control Methods and Use of Equipment Signals

Table 23.1 includes many methods in addition to univariate SPC, such as multivariate SPC and Real-time SPC. In the semiconductor industry, multivariate methods are predominantly used on equipment signals. Because equipment signals are measured during processing, the term "real time" has been associated with SPC on equipment signals. Because many of the methods used with equipment signals are broader than the univariate SPC, the generic term Fault Detection is also generally used for Abnormality Control applied to equipment signals. The sources of most of the equipment signals are regulatory controllers on the equipment. This application of abnormality control to the regulatory controllers is the layer labeled "Monitor Regulatory Controllers" in Figure 23.3. The regulatory controllers will maintain the controlled output signals close to the setpoints specified in the recipe, even under the influence of moderate faults. However, the actuator values required to maintain the controlled outputs to their setpoints in the face of

faults is usually significantly different than from the situation with no faults. Thus, the actual source of real-time data used in the fault detection technique is the actuator values from the regulatory controllers.

The mathematics used in equipment signal monitoring ranges from very simple to quite complex. The reader is referred elsewhere for a discussion of the mathematics [1]. The bibliography at the end is also useful for identifying typical conferences, proceedings, and web sites which also cover this topic more extensively. There is similarity in the methods to those used for monitoring the Run to Run controller (covered in Section 23.8). However, due to the volume of variables and the within run time aspect, real-time controllers are more difficult to monitor.

23.6.4 Definition of the Sensitivity vs. Robustness Challenge

In the section discussing error rates, the concepts of false positive and false negatives were introduced. Let's define sensitivity as the ability to more easily detect errors, i.e., the test is more sensitive to smaller errors. Thus, sensitivity is related to false negatives. The smaller the false negative rate, the more sensitive is the technique. Let's define robustness as the ability for the method to correctly function in a wide variety of expected variation. In other words, robustness is related to the false positive rate. The lower the false positive rate, the more robust is considered the method. Based upon the discussion of error rates, it should now be obvious that sensitivity and robustness are trade-offs. This trade-off is specifically highlighted because of the difficulty in achieving a robust and sensitive method. As already mentioned several times, some methods are more appropriate to a particular situation than other methods and thus will yield better error rates. Consequently, they will provide better sensitivity and robustness.

23.7 Specific Compensation Control Methods

23.7.1 When Are Benefits Realizable?

To ensure successful application of a compensation control technique, three items should be considered:

1. Requirements for compensation control to be feasible
2. Definition of goals/improvements to be achieved
3. Probability of success

One must understand that certain requirements must be met for compensation control to be possible. The most basic requirement is that the system must NOT be Independently Identically Distributed (IID). IID was discussed in the section on univeriate SPC. Note that the errors (or random error) can be IID, which is actually preferable. This requirement is the same as saying that the mean itself shifts and drifts. Table 23.11 lists the remaining requirements which relates to the concepts of controllability, observability,

TABLE 23.11 Requirements for Model-Based Process Control

1. Appropriate situation and predictable
   → Shifts or long-term drifts affect machine productivity
   → Typical process variations can be characterized
   → Changes occur infrequently/slowly enough to get adequate disturbance and process models
   → Currently need period adjustment
2. Controllable
   → Adjustments to actuators/settings affect outputs to be controlled
   → Adjustments to actuators/settings can be made quickly enough
   → Adjustments to actuators/settings can be made with enough precision
3. Observable
   → Data collected can be related to output to be controlled
   → Data are/can be collected frequently enough
   → Measurement dead time is short enough
   → Good signal-to-noise ratio

and predictability. In other words, one can predict future changes based upon current and post data, and one can cause alterations in the system to counter-act the changes predicted to occur so that they are never observed in the output. Regardless of the need for productivity or yield improvement, if these basic requirements are not met, then compensation control is not possible.

A clear understanding by everyone of which goals are to be achieved is required to ensure the specific controller is selected which is likely to achieve the specific goals. Table 23.5 lists a variety of possible goals. Implementers and their managers should discuss which goals are desired before possible algorithms are ever discussed. One goal not listed in Table 23.5 is which the particular process selected for implementing control will enable implementation at subsequent similar process. For example, while the improvements expected for this process may be small, it allows the development of the infrastructure necessary for more complicated, high visibility processes which are expected to realize large improvements.

The probability of success relates to basic systems environment, resource management, and change management. These topics are outside the scope of this chapter. However, their importance in instituting a new methodology which is predicated upon automation and interaction with the current CIM environment cannot be stressed enough.

23.7.2 Controller Goals: Tracking the Target, Rejecting Disturbances, and Ignoring Noise

While the final goal of implementing control is usually one of the benefits listed in Table 23.5, the controller itself usually has a particular goal and it is fine-tuned to meet this goal. In other words, particular controller mathematics is used which optimize a particular controller goal in order to achieve an over-all process improvement goal. At first thought, it would seem that tracking the target is the only goal of a controller. However, random noise is also present. The three classic goals of any controller are given in Table 23.12. Trade-offs are required between these goals. Controllers with fast response and which aggressively attempt to remove the influence of disturbances and process dynamics are more susceptible to noise. This susceptibility is due to the amount of data required to distinguish between a real change in the mean caused by a disturbance or process dynamics and a data point in the tails of the normal process variation. Thus, just like for abnormality control methods, Type I and II errors are also of interest in compensation control methods, although rarely discussed in a typical control textbook. In the compensation case, Type I error can be viewed as the controller making an unnecessary action which increases the variance around the target. Type II error can be viewed as the controller not taking a needed action that would have reduced the variance around the target. As the discussion on Type I and II errors mentioned, increasing the sample size requires a greater time for data collection. Thus, a controller with a given algorithm which reduces its type I error, without changing its' Type II error, will be more sluggish to respond. Just as different fault detection algorithms provide better Type I and II errors for a given situational application, so do different compensation algorithms provide better target tracking, disturbance rejection, and noise insusceptibility. We will discuss this concept again in the section on SPC-based compensation control methods.

In traditional petrochemical industry control, stability can be a real issue. In fact, in these industries, stability considerations may dominate the design and tuning of the controller. However, in run to run

TABLE 23.12 The Three Classic Goals of a Controller

1. Track the target without lag
   → The output should immediately go to the new target (set point) when it is changed
   → The output should remain on target even under the influence of process dynamics
2. Prevent disturbances from influencing the output
   → The output should remain on target even when disturbances occur
3. Ignore random noise
   → The controller should not increase product variance by responding to spurious (not real) fluctuations

control, it is rarely a large concern. Thus, this lack of focus on stability concerns has also led to a difference in the approach to control between the semiconductor industry and the petrochemical industry.

23.7.3 Feedback/Feedforward Control

All compensation control techniques can be classified as Feedback or Feedforward. Note that Table 23.2 list Feedback and Feedforward control. Figure 23.8 demonstrates these two types of controllers. Feedback uses measurements about the current process results to decide how to change the process for the next sampling period. Feedforward uses measurements on incoming materials, process, or equipment to decide how to change the process for the current process. Thus, feedback control drives the average value to target (i.e., drives C_{pk} to equal C_p). Feedforward control, because it accounts for incoming variations, can improve the C_p value, i.e., it turns apparently random variation into non-random variation, which can be compensated for. Based on Table 23.6, feedback control is used to compensate for expected disturbances. Feedforward control is used to compensate for measured disturbances that have been modeled. Target (Setpoint) changes are encountered in both feedback and feedforward control.

23.7.4 Common Compensation Control Methods Used for Run-to-Run Control

There are a variety of control methods in the control literature, although many are not currently deployed in the Semiconductor Industry. In this section, how run-to-run (batch-to-batch) control evolved in the semiconductor industry will be covered. Control methods from the petro-chemical industries will be introduced next. Finally, the critical issue of deadtime will be introduced. This section will conclude with an example that illustrates many of the concepts introduced in this section.

23.7.4.1 The Creation of SPC-Based Controllers in the Semiconductor Industry

As mentioned previously, Statistical Process Control, an abnormality detection and control method, has been practiced for many years in semiconductor manufacturing. However, in several cases, traditional SPC just would not, and could not, work [15,30]. The assumption that the expected variation can be described by IID, i.e., that the values were uncorrelated, did not hold. Because of the variety of processes run on a single piece of equipment, a given process produced a small sample size causing statistical problems. In addition, the assumption that the machine was "out-of-control" when an SPC test alarmed was not correct. To many, "out-of-control" meant the machine needed to be repaired, and yet it was not broken. Others interpret the philosophy of using a "knob" to re-turn the process when an alarm is given as compatible with SPC concepts [31,32]. However, the alarm frequency was greater than the traditional

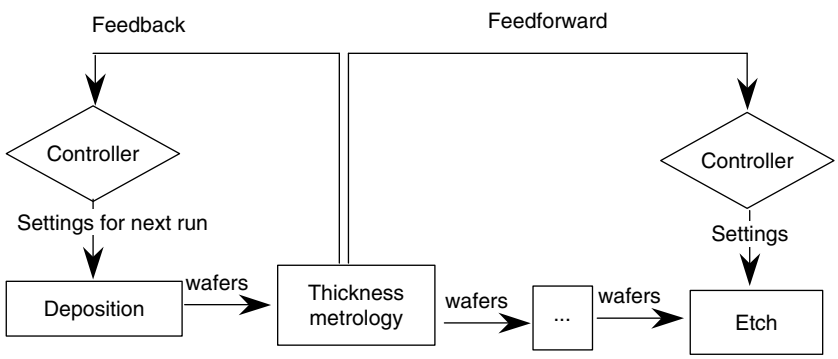


FIGURE 23.8 Feedback and feedforward controls.

SPC expectations, i.e., traditional SPC expects alarms to be infrequent and an exception to typical behavior. Thus, because of the frequency of the knob adjustment, ensuring consistency in how different operators used the knobs and documenting what values they used is extremely important. In the mid-1980s, the search began for a method to determine:

- When to re-adjust the process
- How much to re-adjust the process

In addition, the emphasis was on:

- Ease of implementation due to the frequency of the adjustments
- Prevention of over-controlling due to the inventor's SPC background
- Ability to use the methodology in an environment of small number of runs per process

Thus, the SPC chart gained a new role: to determine when to adjust the process. By allowing the SPC chart to trigger when to adjust the process, over-control could be avoided. It was also believed to be easy to implement since the fab already had SPC charts in place. In addition, the operators and engineers were already used to taking an action when a SPC failure occurred. However, a traditional SPC chart would not be appropriate since a controlled output would be serially correlated, which violates SPC chart assumptions. Thus, a regression control chart is used [33], on which the residuals (Actual value–Predicted value) are plotted. If the predictions are good, then the residuals should be IIDN $(0,\sigma)$ and a SPC chart is appropriate. It is assumed that a model is used for the predictions. When the model is no longer centered the predictions will become poor, causing a fault to trigger on the chart indicating the model needs to be tuned. Of course, in such a case, the residuals may also become autocorrelated, i.e., the residuals will not be independent (the second I in IIDN). Due to the autocorrelation, the error rates (α and β) will not equal the values expected from the SPC chart set-up on truly IID$N(0,\sigma)$ data. However, in practice, tuning based upon an SPC chart has worked as expected by the engineers who were used to operators manually tuning based on SPC chart on the controlled output.

To handle the small run sizes, it was envisioned that models could be used to combine data from different processes into one controller. Thus, if different processes were used for different thicknesses, a model was used for thickness to combine all the data into one model:

$$\hat{y} = mx + b \tag{23.10}$$

where $\hat{y}$ is the model prediction for output y, e.g., thickness; m the slope of the model (model parameter), model coefficient, gain; x the recipe setting, e.g., time; b the offset and is the tuned model parameter; and y the output being modeled, e.g., thickness.

The next question was what data to use to tune the process. An obvious answer was to use the data involved in the WECO failure [30,31]. To show how the process was tuned, b was tuned with the data in the WECO violation set by

$$\text{bnew} = \text{bold} + \frac{\left\{\sum_{i=1}^{\text{NW}}(y_i - \hat{y}_i)\right\}}{\text{NW}} \tag{23.11}$$

where NW is the number of runs involved in WECO test failure, e.g., the 4 or 5 runs involved in a 4/5 WECO run rule failure, called the violation set.

Every time the model is tuned, the history is re-set. In other words, the historical data used to calculate the next SPC test does not include any data previous to or including the run for which the last SPC trigger occurred. In addition, the history is re-set after maintenance. If the process is expected to return to a baseline state, the b value is re-set to its original baseline value. However, because the process may act

significantly different after each maintenance, test runs may be used to calculate the b value:

$$\text{bnew} = \text{bold} + \frac{\left\{\sum_{i=1}^{\text{L}}(y_i - \hat{y}_i)\right\}}{\text{L}} \tag{23.12}$$

where L is the number of test runs, usually 1, bold the last value of b or the baseline value of b.

To calculate the setting (x) to be used in the recipe, the value of x for the next run was found by solving Equation 23.10 by setting the prediction equal to the desired target value T and using the new b value for Equation 23.11 or Equation 23.12:

$$\hat{y} = T = mx + b \rightarrow x = \frac{T - b}{m} \tag{23.13}$$

where b is the most up-to-date value of b (from Equation 23.11 or best estimate) and T, the desired target value for output y.

The above methodology became known as "model-based process control" in the semiconductor industry because it deployed a process model as part of the controller. A simple graphical representation is shown in Figure 23.9. Table 23.13 lists the algorithm with some generalizations that will be made clear in later sections. Representative equations to use in the algorithm is shown as sub-bullets in Table 23.13. It is the tuning of the model that provides feedback, i.e., the above algorithm is a form of feedback control. Control for setpoint changes occurs when the value for T is changed in Equation 23.13 resulting in new settings values. While Figure 23.9 shows the idealized case of a linear process, no noise, and no model mismatch, experience has proven that the method works even in the presence of these complications. Model mismatch is when the coefficients of the model other than the offset, i.e., the gains are not equal to the true values of the process or when terms, such as a quadratic term, are missing in the model.

To use this method, a model needs to be fitted to the process data to obtain a value for m and an initial estimate for b in Equation 23.10. Sometimes the equations used are not of the form above, and it may appear that a model of the form of 10 is missing. However, the same mathematics is implied, i.e., the equations used can be put in the form above. For linear single input-single output systems, Equation 23.10 seems unnecessary because $\hat{y} = T$ can usually be substituted in Equation 23.11 and Equation 23.12 and an algebraic solution is possible, as shown in Equation 23.13. However, using an explicit model form is a good practice because such a form is necessary for non-linear, multivariable systems, as will be discussed later.

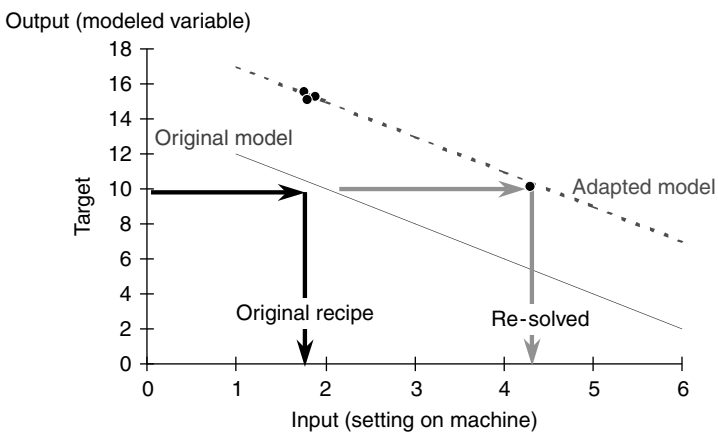


FIGURE 23.9 Graphical representation of model-based process control for case of low error, linear process.

TABLE 23.13 Generalized Semiconductor-Model-Based Run-to-Run Control Algorithm

1. Model created to predict output as a function of input (process model)
   - Equation 23.10
   - Equation 23.25
   - Others

   → In Figure 23.9, the lower line is the original (untuned) model
2. Model is solved (optimized, inverted) to determine what value of the settings is predicted to give values for the output equal to target

   Note: While settings are calculated each run, they are not necessarily different each time
   - Equation 23.13
   - Equation 23.27
   - Others

   → In Figure 23.9, target value of 10, $X = 1.8$
3. Wafers are run with these settings values in the recipe
4. Measurements are made for run
5. Output metric (actual value) is calculated from measurements
   - For example, average thickness = average (measured values)

   → In Figure 23.9, actual output is higher than target
6. Compare actual output to predictions to decide whether the model should be tuned; if not, skip to step 7
   - Statistical process control chart
   - Deadband, Equation 23.24
   - Others

   → In Figure 23.9, is determined the need to tune
7. Tune model (provides feedback) according to disturbance model
   - Equation 23.11 if production run
   - Equation 23.12 if qual run
   - Equation 23.22 or Equation 23.23 with λ value for production or qual run
   - Others

   → In Figure 23.9, new tuned model shown (adapted model, dotted line)
8. Go to step 2

   → In Figure 23.9, new solution results in output on target

With example equations and relation to Figure 23.9.

23.7.4.1.1 Process Model and Disturbance Model

Equation 23.10 through Equation 23.12 can also be reformulated to highlight two important concepts.

- The process model which predicts how a change in the input will cause a change in the output
- The disturbance model that predicts how the output changes over time (runs)

Using the concept of the process model and the disturbance model, an equivalent equation can be presented.

$$\hat{y} = \text{Process model} + \text{Disturbance model} \tag{23.14}$$

Thus, by comparing Equation 23.14 with Equation 23.10, Equation 23.11, and Equation 23.12, it's obvious that:

$$\begin{aligned} mx &= \text{Process model} \\ \text{How } b \text{ is tuned} &= \text{Disturbance model} \\ &= \text{Equation 23.11 and Equation 23.12} \end{aligned} \tag{23.15}$$

It is the formulation of Equation 23.14 that allows expansion into other non-linear and multivariable process models and more advanced disturbance models. It also explicitly highlights that one must effectively capture the inherent behavior of the system to input changes (the process model) AND be able

to predict the way disturbances will impact the system (the disturbance model) to be able to effectively control the process. Thus, Equation 23.14 demonstrates how feedback control is a mechanism for compensating for the disturbance.

23.7.4.1.2 The Biggest Benefits of "Model-Based Process Control"

Several major benefits come out of the above methodology and are listed included in the list of Table 23.5. Models were originally used because processing was needed at more than one desired target, due to different products needing different results or due to the same machine being used for two different processes in the flow. Thus, a model-based methodology was created to allow one control strategy for both processes. Once the controller could handle two processes, it could handle infinitely many [30] providing increased process flexibility. In addition, since all the data were related through the model, each product run became the "test run" for the next product run, thereby reducing the need for look-ahead and qualifications runs which decreased pilot usage too. This ability to model more than one process is the "global modeling" concept introduced in the section on Characterizing Control Needs.

With the automation of the above methodology, engineering, technician, and operator productivity was improved. By allowing the controller to decide when qualifications were necessary based upon the data and using the controller to perform quals, including those for post-maintenance recovery, qual runs, qual times, and post-maintenance recovery time were reduced. Using the controller data to determine cause of alarms speeds the repair time. The models can be unique per machine chamber and the controller used to drive chamber results to sameness removing the effect of chamber/hardware differences. Other benefits listed in Table 23.5 will be discussed in the section on controller monitoring.

The controller described above was applied to the Plasma Enhanced Chemical Vapor Deposition (PECVD) of tetraethyl orthosilicate (TEOS) [15]. The C_{pk} value was increased from 2.5 to 4.5. The number of quals was reduced from one per day to one every 3 days. Due to one pilot wafer was used for each qual, reducing quals also reduced the associated number of pilots. The time spent doing quals was reduced from 1 h per qual to 15 min per qual. Other examples of improvements are given in Ref. [15].

One item to note is that the SPC based methodology appears to be so simple as to not require automation to achieve the benefits cited. However, in practice, the necessity of automating the calculations to achieve adequate quality and consistency has been demonstrated. On the other hand, it has been found that entering the new settings into the equipment does not necessarily require automation, although increased quality will be achieved with automatic recipe download to the equipment.

23.7.4.1.3 The Biggest Issues with SPC Based Process Control

One issue with WECO based tuning is the resulting "bang–bang" nature of the output. Bang–bang control is the simple style of control used for home thermostats. The setting can take on only two values, on or off. The result is the temperature will alternate between above or below the desired target temperature, but rarely at the target temperature [4]. Similarly, with the WECO tuning, the output generally alternates between above and below the target. Contributing to this behavior is the small sample of data used in the tuning. Additionally, lag can sometimes be encountered due to waiting for a SPC failure. The issue of type I and II errors needed to be addressed differently when the response action is an automated control action verses. a manual control action. Table 23.14 compares the cost of type I and II errors for the case of expected variation needing compensation only, i.e., the possibility of the system encountering unexpected variation is ignored for this discussion. As can be seen from this table, the trade-off between Type I and II errors should be different for automated and manual controls. In the case of automated control, minimizing Type II errors has a heavier focus due to the smaller cost associated with Type I errors. Obviously, this trade-off must be within reason or the overall cost will not be favorable due to considerable overcontrolling. Finally as Equation 23.14 highlights, one must have a good disturbance model to effectively compensate for the disturbances. In many cases, the SPC based tuner

TABLE 23.14 Comparison of Error Costs for Manual vs. Automated Control for Case of Expected Variation Needing Compensation

Type	Description	Costs in Manual Case	Costs in Automated Case
I	Take control action when none is needed	Wasted time trying to track down fault that does not exist: unnecessary machine downtime; if action finally taken, will result in overcontrol (increased variation around target) and will need to manually undo action in future	Control action easily taken by system: action will result in some increase in variance around target; action will need to be undone, but control system will detect overcontrol within next few runs and undo action itself
II	Note take control action when one is needed	Loss of quality	Loss of quality

does not adequately represent the disturbances dynamic behavior. As noted several times, when the mathematics more adequately represents reality, Type I and II error rates are generally improved.

23.7.4.2 The Creation of EWMA-Based Controllers in the Semiconductor Industry

As the search continued for a method with better Type I and Type II trade-offs, many tried the exponentially weighted moving average (EWMA) to tune the b value of Equation 23.11 [34–39]. The EWMA has the distinction of being known as an EWMA chart to SPC experts [40], as a first older digital filter to control pracitioners [4], and as integrated moving average time series, represented as (1,0,1), to statisticians [41]. It's invention in three different fields attests to its natural applicability to many processes, i.e., it was expected that an EWMA would be a good representation of the disturbance dynamics.

The equivalency of the first-order digital low-pass filter to an EWMA can be seen by the EWMA equation

$$Z_t = \lambda^* V_t + (1-\lambda)^* Z_{t-1} \tag{23.16}$$

where Z_t is the filtered value at time t, V_t the variable to be filtered, λ the filter factor, 1/time filter time constant.

Note: Whether the filter factor is on the output variable or on the filtered variable varies article to article and software to software. This discrepancy is responsible for many an engineer spending wasted hours trying to track down a math error that is in fact is a misunderstanding in filter factor usage.

To use the filter for prediction, the following equivalence is used:

$$\hat{V}_{t+1} = Z_t \tag{23.17}$$

where $\hat{V}_{t+1}$ is the prediction of the variable V for the next run (time$=t+1$), Z_t the filtered value of variable V for the current run (time$=t$).

Another way to represent the EWMA that more clearly shows how the EWMA corrects itself using a fraction of the prediction error is to rearrange Equation 23.16 and note that Equation 23.17 says the prediction for variable V at time t is the filtered value Z at time $t-1$

$$Z_t = Z_{t-1} + \lambda^*(V_t - Z_{t-1}) = Z_{t-1} + \lambda^* e_t \tag{23.18}$$

where e_t is the error in predicting $V_t = V_t - Z_{t-1}$.

The EWMA is used for control by filtering the model offset b, i.e., defining

$$V_t = b = y - mx \tag{23.19}$$

$$\text{bnew} = Z_t \tag{23.20}$$

$$\text{bold} = Z_{t-1} \tag{23.21}$$

and substituting into Equation 23.16

$$\text{bnew} = \lambda^*(y - mx) + (1-\lambda)^*\text{bold} \tag{23.22}$$

or equivalently, substituting Equation 23.19 through Equation 23.21 into Equation 23.18

$$\text{bnew} = \text{bold} + \lambda^*(y - \hat{y}) \tag{23.23}$$

The value for λ is determined by what will result in the best control for the system with its associated control issues. A different value of λ may be used for production wafers and qual runs to mimic the concept in Equation 23.11 and Equation 23.12. For many cases, since only one qual is run, λ for quals is 1, i.e.,

$$\lambda_{\text{production}} = \lambda$$

$$\lambda_{\text{qual}} = 1$$

The EWMA-based controller has found wide applicability in the semiconductor industry [34–39]. In some cases, the tuning is for every run [34,36–39]. In other cases, WHEN to tune is still decided by a SPC chart [35,39] or a deadband. A deadband is used for determining when to tune by

$$\begin{gathered}\text{If} |\text{bnew} - \text{bold}| \leq \text{db, then bnew} = \text{bold} \\ \text{Otherwise, bnew} = \text{calculated bnew value}\end{gathered} \tag{23.24}$$

where db is the deadband value.

The deadbanding prevents unnecessary control action similar to the SPC chart without suffering from unnecessary lag. It is another way to achieve type I vs. type II trade-off. It can be considered equivalent to the trigger value in an EWMA chart. It is also used to prevent what the engineer considered undesired frequent small control action, especially if the changes are being made manually to the recipe on the machine. The deadbanding may be implemented in other ways, such as checking if the input (x) would change greater than a certain amount. If not, then tuning will not occur.

23.7.4.3 Extension to Multivariable, Non-Linear, Constrained Systems with Feedforward Control

Equation 23.10 and Equation 23.13 assume linearity and single input-single output, as well as lack constraints on the inputs or outputs. The method above can still be used by changing Equation 23.10 to

$$\hat{Y} = f(X, \text{FF}, \hat{Y}, \theta) + b \tag{23.25}$$

where $f(x)$ is the generic equation, including non-linearities; X denotes more than one manipulated variable; FF denotes the feedforward measured variables; Y denotes more than one output variable (actual value); $\hat{Y}$ denotes prediction for each output variable; θ denotes all the parameters involved in equation f; and b the current estimated value of the disturbance, i.e., $b = \text{filtered}\{Y - f(X, \text{FF}, \hat{Y}, \theta)\}$.

In reality, Equation 23.25 is actually several equations, one for each output. The effect of a feedforward variable is easily seen since it is included in the model. For example, an incoming thickness value could be used as a feedforward variable for a controller on final thickness using time as the manipulated variable.

$$\text{Final thickness} = \text{initial thickness} + m^* \text{time} + b \tag{23.26}$$

Based upon the discussion of Equation 23.14

$$f(X) = \text{process model}$$

and

$$\text{HOW } b \text{ is tuned} = \text{disturbance model.}$$

The biggest change due to the non-linearities and the constraints is the need for a numerical solver to replace Equation 23.13. In such a case, there are several cost functions that can be minimized. A common cost function that allows trade-offs between changes in the manipulated variables (size of process adjustment) and reaching the target is

$$\underset{X}{\text{Minimize}} \left\{ \sum_{i=1}^{N} W_i (T_i - Y_i)^2 + \sum_{k=1}^{k} w_k (X_k - X_{k,t-1})^2 \right\} \tag{23.27}$$

with constraints

$$g_j(X) \geq 0$$

$$h_{\text{L}}(Y) \geq 0$$

Constraints are used because not all values of X are allowable. Constraints on the outputs are used because all outputs generally have upper and/or lower specifications and some outputs may not have targets, but rather just upper or lower constraints with no losses associated with being anywhere inside the specification, only with outside the specification. Additional procedures may be sued to prevent oscillations in the solution [16].

The changes to the linear univariate algorithm can easily be understood by examining Table 23.13. The generalization of the algorithm to use alternative tuning methods, solution methods, and process models is apparent. By the inclusion of feedforward variables, feedforward control in additional to feedback control is achieved. If the model is not tuned, i.e., the desired control output is not measured, then only feedforward control would result.

23.7.4.4 Transformations to Comply with Additive Disturbance Assumption

Equation 23.14 implies the assumption that the disturbance is *additive* to the modeled output. This assumption puts constraints on the output used for to modeling. However, this assumption has been shown to be valid for a variety of systems. Thus, the output modeled may be different than the output controlled in order to meet the assumption.

For example, let thickness be the desired output to control. However, the data shows that the additive disturbance assumption really applies to rate, i.e., it is rate that is changing. In this case, Equation 23.10 is replaced by

$$\text{Râte} = b \tag{23.28}$$

where rate, the modeled tuned output; râte, the prediction of output rate; b, the tuned offset, tuned using SPC-based methods, or EWMA, etc.

Instead of solving Equation 23.10, as described in Equation 23.13, a new equation is created which is solved.

$$\hat{\text{Thickness}} = \text{Râte}^{*}\text{Time} \rightarrow \text{Time} = \frac{\text{Râte}}{T} \tag{23.29}$$

where thickness, the controlled output; $\hat{\text{Thickness}}$, the prediction of controlled output; T, the target for controlled output (thickness); time, the setting in recipe used to adjust process, and râte the use latest tuned value.

23.7.4.5 Predictor Corrector Control

Sometimes, the EWMA model is not an adequate representation of the disturbance. This situation occurs when the disturbance is a near-constant drift run-to-run. The EWMA controller will result in constant offset. (This situation will be examined further in a later section on control of metal sputter deposition.) Thus, a way to correct the prediction of the EWMA controller was needed [16,42]. The resulting disturbance model was called the predictor corrector control (PCC) and remembering the discussions going with Equation 23.16, Equation 23.17, Equation 23.19 through Equation 23.23, and Equation 23.25

$$\begin{aligned} S_t &= \lambda^{*} V_t + (1-\lambda)^{*} Z_{t-1} \\ \hat{T}_t &= \beta^{*} T_t + (1-\beta)^{*} \hat{T}_{t-1} \\ T_t &= V_t - S_{t-1} \\ Z_t &= S_t + \hat{T}_t = \text{Filtered(smoothed) value} + \text{Filtered trend} \end{aligned} \tag{23.30}$$

where Z_t is the filtered value of variable V at time t, V_t the variable to be filtered ($= y - mx$ in the simplest case) $= b$, T_t the current trend (i.e., the difference between the current value and the previous filtered value), S_t the current smoothed value, and $\hat{T}_t$ the current smoothed trend.

As can be seen, Equation 23.30 assumes a constant trend and the purpose is to estimate not only the current value of the model offset, but also the trend (or rate of change) in the model offset.

23.7.4.6 Comparison to Other Methods

To demonstrate that an EWMA controller is equivalent to a pure integral controller, which is frequently found in use as a real-time equipment controller, or that PCC is similar to a double integrator, is beyond the scope of this chapter. However, the reader should note that to facilitate such comparisons, different equation forms derived above (such as Equation 23.22 vs. Equation 23.23) were developed to allow for easier comparison to some modern control methods. See elsewhere for a comparison to modern control techniques and the relationship to stability [16,17,42,43]. The typical semiconductor control technique may also be considered a deadbeat controller in that the set point is not filtered [4]. However, because run-to-run control is not truly a digital controller on a continuous process, the analogy is not quite exact. However, with time some readers may find a discussion of deadbeat controller helpful and they are referred elsewhere [4]. Others may find reading on real-time optimization (RTO; NOT rapid thermal oxidation) in the process industries fascinating [44]. The RTO is very similar to multivariable run-to-run control but is applied to a continuous process. Algorithmic SPC's compensation method [5,6], which was discussed previously, is very similar to the minimum variance control of MacGregor [45]. Minimum variance control (as well as internal model control) are equivalent to the generalized semiconductor multivariable controller presented above. However, Algorithmic SPC adds an additional term in the disturbance model to account for measurement noise, which they call an extrinsic error. The main resulting difference in the controller design will be different values in the correlated noise parameters.

23.7.4.7 Treatment of Dead Time (Measurement Lag)

Measurement lag is the time between the run and the controller getting the measurement. It can have very detrimental effects on the control performance. There are some existing methods for dealing with lag, although they generally assume CONSTANT lag. One method is discussed in describing the compensation method of algorithmic SPC [5], whereby they alter the best forecast for the model error term. They stress that this alternative equation must be used always should the possibility of deadtime occur since the optimal results will not be obtained otherwise. They also note the difficulty of varying deadtime, even if the deadtime is known. However, in general, while the predictions can be modified to account for the deadtime, the controller performance may be too degraded to be acceptable. Thus, one must consider measurement lag and in some cases, eliminate it, to achieve acceptable controller performance.

23.7.4.8 Summary Example Using Metal Sputter Deposition

To illustrate many of the concepts presented above, especially of the concept of the additive process and disturbance models, metal sputter deposition will be used as an example [46]. The metal to be deposited, titanium, had thickness goals spanning 400 Å due to the wide variety of processes used. Before the controller was installed many test runs and monitor pilots were used. The C_{pk} value was near acceptable, but the technician and processing overhead was too expensive. Note in this example, target means the desired value for the output (thickness), not a sputter target.

First, a process model needed to be created which spanned all the processes. Figure 23.10 demonstrates a model where the output was rate, rather than the output that had a goal (thickness), and the input was thickness, rather than time. The "hidden" model is that thickness = rate × time. The model appears to capture the process data eloquently, but would it allow the assumption of the *additive* disturbance model? Figure 23.11 shows that over a long time period, the model moved up and down, i.e., the model offset varied, but not the slope. Therefore, the additive disturbance model is correct. Note that what is not shown is the other attempts at process models. Within a short period of time, the control engineer learns that only a couple of functional forms will represent a wide variety of processes from lithography to sputter deposition. Thus, once these forms are known to the engineer, they merely attempt to match the data to one of these forms. Consequently, process model building is faster than one might expect at first glance.

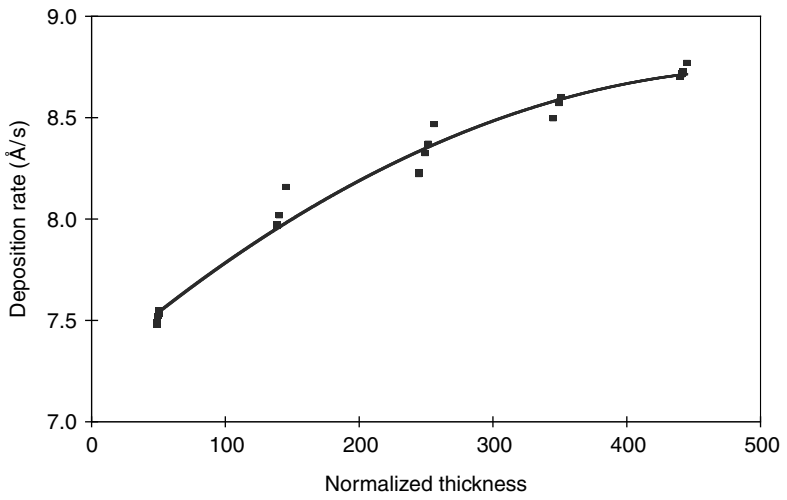


FIGURE 23.10 Process model for Ti sputter deposition.

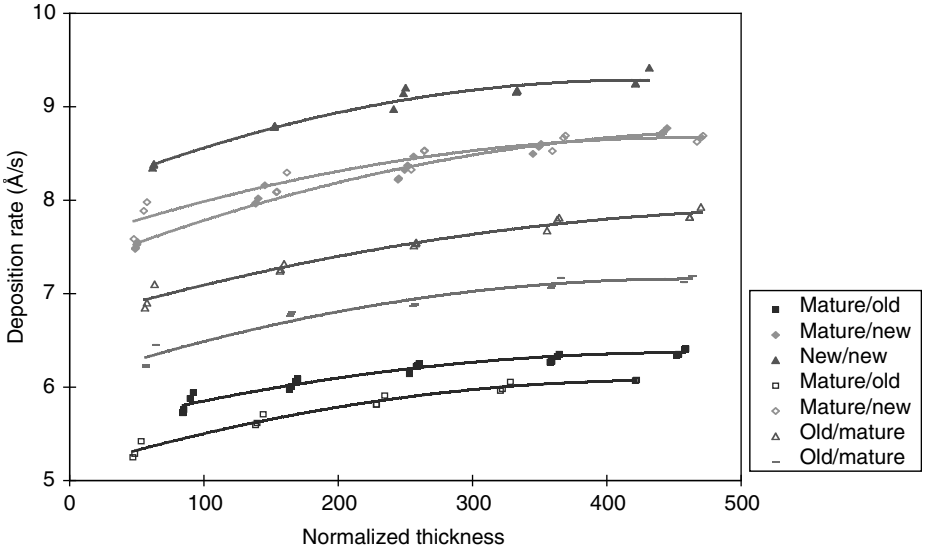


FIGURE 23.11 Disturbance model: if process model in correct space see offset only over large time periods.

Next, the disturbance model must be created. First, an EWMA was attempted. The results are shown in Figure 23.12. An offset appears to be presented, which is more obvious in Figure 23.13. Note how the offset is near constant. Based on Figure 23.12 and Figure 23.13, the PCC model was attempted next. The results are shown in Figure 23.14 and Figure 23.15. Now, the offset is gone, and the offset is centered around 0. The final results were test runs were eliminated, reduced the number of monitor wafers by

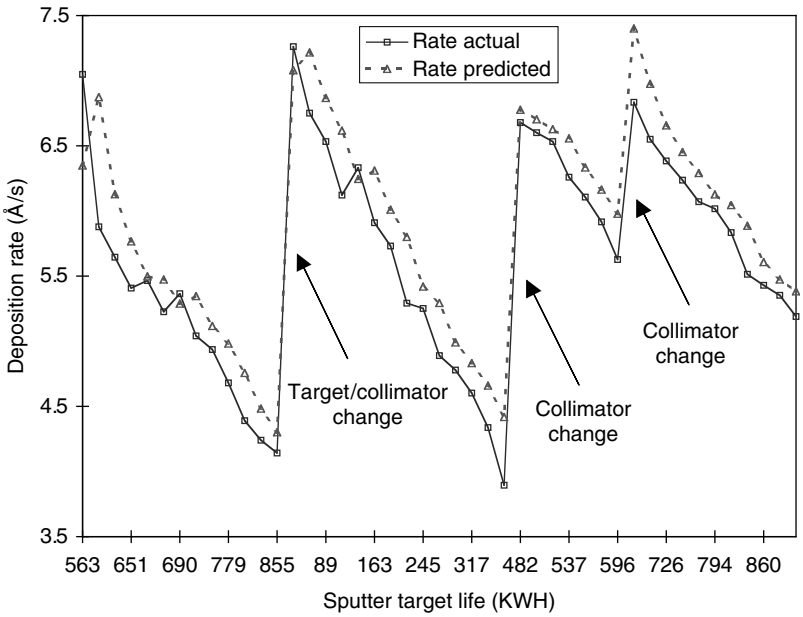


FIGURE 23.12 Incorrect disturbance model.

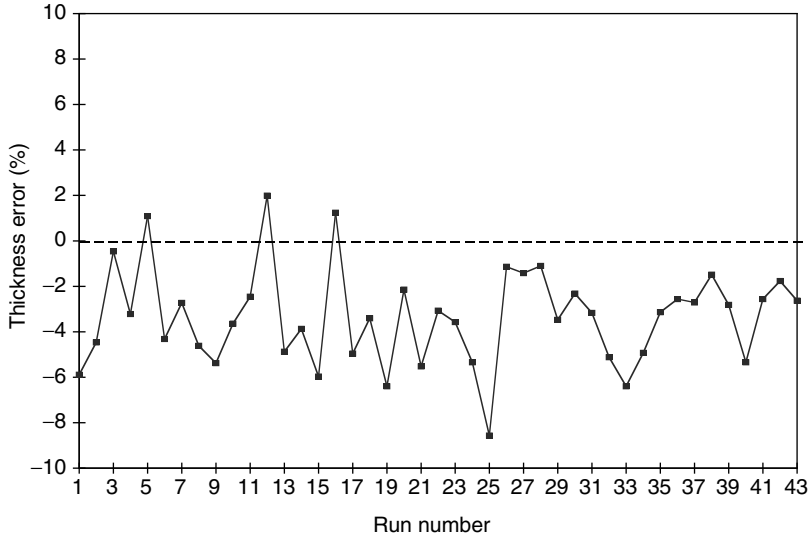


FIGURE 23.13 Offset easier to see in actual target (thickness is variable with target).

a factor 3, C_{pk} improved by 10%, and simplified processing for operators and reduced sustaining effort for engineers

23.7.5 Real-Time Compensation Control Methods

Table 23.15 provides a list of new activities in real-time equipment control. This table does NOT mention pressure, MFC, or radio frequency (RF) controllers on etchers. While these are extremely common

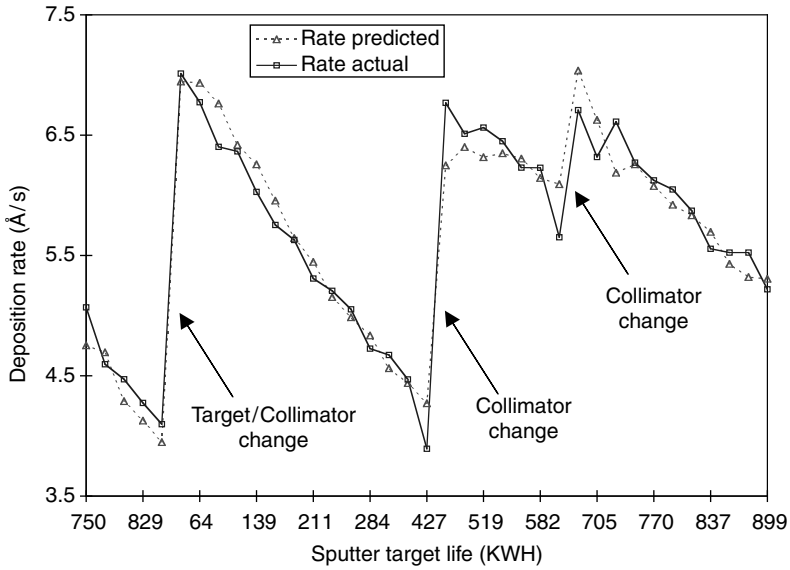


FIGURE 23.14 Better disturbance model: linear drift (predictor corrector control).

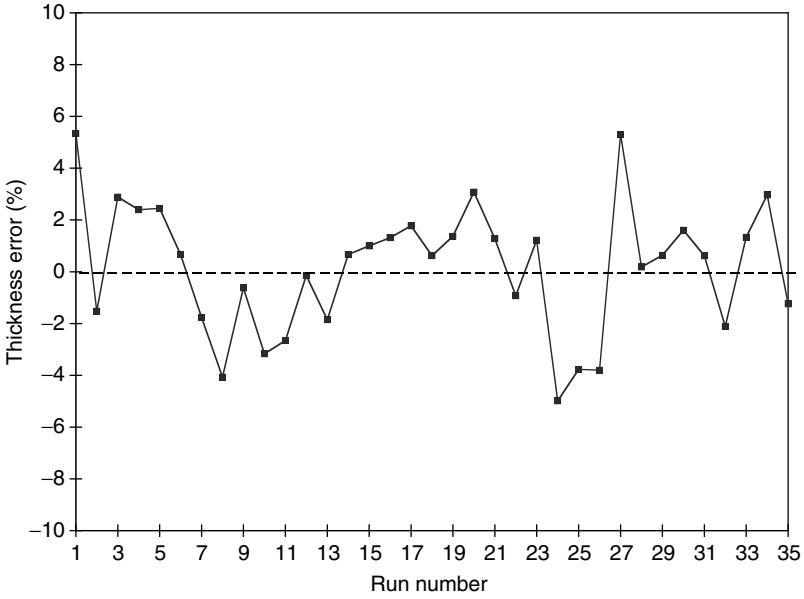


FIGURE 23.15 Easier to see improvement in actual target (for thickness variable).

controllers, significant changes have not occurred in their control ability. Mass flow controllers are now becoming digital, but their basic algorithms are not changing. Changes are expected in the RF controller arena, but right now, most changes are due to RF sensor installation rather than RF control.

Table 23.15 also does not cover control activities in rapid thermal processing (RTP). Chapter 11 covers RTP. In that chapter, is a section on RTP temperature control systems. The reader is referred to Chapter 11.

23.8 Monitoring the Supervisory Run-to-Run Controller and the Controller System Advanced Process Control

As was mentioned when APC was introduced, APC includes abnormality methods for monitoring of the compensation controller. Thus, some concepts of type I and type II errors for a control system will now be introduced. Next algorithms for detecting abnormalities in the control system are introduced. It is amazing how these extremely simple algorithms can prevent horrendous yield loss problems due to the increased sensitivity to faults they provide.

TABLE 23.15 New Activities in Industry in Real-Time Control

- Uniformity control
  - Rapid thermal process
  - Lithography bake plate
- Model predictive control on furnaces
  - Cycle time and performance improvement
- Endpointing
  - Small open area
- Sensors
  - Purchased integrated from process vendor, manufactured by sensor vendor

23.8.1 The Other Type I, Type II Errors: Detection of Change in Overall System

There is another error situation present in compensation control methods. This situation is when the controller operates on a system with different variances than for which the controller was designed. In other words, the random and/or non-random variances are different than expected. Although the controller may still drive the output to target, the overall quality of the result is suspect since operating in this regime has never been qualified. Thus, unmeasured variables and the metrology itself may be out of specification. The desired behavior would be for the controller to detect the change in behavior and generate an alarm. One can consider Type I error as the controller generating an alarm when the system variance has not changed, and Type II error as the controller not generating an alarm when the variation is different.

Note that this situation can also arise in the abnormality based methods when the output appears to be on target, but in reality the system has drifted to a state where the metrology and the system are out of specification. In this case, abnormality based methods have no extra ability to detect the system change. Fortuitously, the compensation behavior provides another way to determine if the system is behaving as expected. In other words, the compensation based controller can monitor not only the output changes, but how the output changes in response to a given change in input as well, to determine if the over-all system has changed behavior resulting in a better overall Type I and II errors. This desire to monitor for abnormalities in the compensation based controller leads to the concept of merging both methods into one, one of the conceepts of Advanced Process Control.

23.8.2 Methods for Monitoring the Supervisory Controller

Monitoring the various aspects of the compensation controller would appear difficult. However, monitoring the feedback amount in a model based control system simplifies the abnormality detection. The simple monitoring algorithms are:

- Determine if the adjustment is unusual (single or cumulative adjustment too large, too frequent)
- Determine whether the output responds to the adjustment as expected, including random noise level

A product called Overseer [30] was created to assist with the automatic checking of the Supervisory Controller. It added some more intricate tests, such as:

- Are all equipments of a given type drifting or shifting in unison

The main concept to be left with the reader is that no controller should ever be implemented without sufficient monitoring. The monitoring prevents the controller from driving the system past reasonable operation, while the control improves the sensitivity and robustness of the abnormality checking.

23.9 Continuous Process Improvement

The description of the various control methods will end with a discussion of the outer control layer in Figure 23.3, continuous process improvement. At all times, engineers are examining their process to determine how best to increase it's productivity and yield. They employ engineering knowledge, statistics, and common sense. However, it is the data coming from the various levels of the controllers, in conjunction with data from specially designed experiments, which provides the most information about how best to improve the process or equipment.

23.9.1 Benefits of Reducing the *Effective* Noise of the System

As mentioned in the introduction on types of errors, the type I (α) and type II (β) errors are a function of the sample size, the particular size of change (Δ) to be detected, and the noise level of the system (σ). For a fixed sample size, both the type I and type II errors decrease as the noise (σ) decreases. Conversely, because β is the probability of not detecting a difference when there is one, decreasing the noise also increases the probability of detecting a difference. The probability of being able to detect a change is called the power and is equal to $1-\beta$. Having good power is necessary to evaluate process improvement changes and design new products [47]. By using global modeling and feedforward variables (including wafer order), the *effective* noise of the system is decreased, and, consequently, the power to detect a change is increased. Thus, the procedures outlined in the section on Abnormality Detection Control Methods to enable robust fault detection are also important to continuous process improvement.

23.9.2 Comparing How Different Machines React

Investigations of why different machines of the same model number respond differently to the same inputs provides useful information that can be exploited to improve the machine. Determining that one machine is a rogue tool based upon its response to input changes generally provides increased sensitivity to a machine likely to be the source of yield loss at some point. Thus, controller data provides a great opportunity to identify rogue machines. Driving the machines to respond to inputs in a similar fashion is the greatest assurance that the outputs will be consistently similar for all equipment, a major desire by the fab's customers.

23.10 Summary

This chapter conveyed the depth that process control in the semiconductor industry has achieved. The reader should now be able to assess the variety of control tools available and select those most appropriate for application in their facility. As part of the references section, several pointers to web pages and further resources are provided.

23.11 Acronyms and Glossary

Abnormality control methods: Methods based upon detecting abnormalities and correcting them.
Alpha (α): Percentage of false positives (detecting an abnormality when one has not occurred).
Advanced process control (APC): A combination of abnormality and compensation control methods. Also equal to FDC plus model-based process control.
Beta (β): Percentage of false negative (not detecting an abnormality when one has occurred).
CIM: Computer integrated manufacturing (may imply MES).
Compensation (target tracking) control methods: Methods based upon actively compensating for expected sources of variation.
C_p, C_{pk}: Most common process capability indices used to assess the process' ability to achieve yield. C_{pk} considers how centered the process is, while C_p does not.
Disturbance model: A model that predicts the way disturbances will impact the process, i.e., describes how the process ages.
Dynamics: The non-random behavior of the system over time, i.e., how the output would change with each run if no compensation were used.
Error rate: Rate at which a false positive error {alpha (α)} or false negative {beta (β)} occurs.
EWMA: Exponentially weighted moving average (filter, type of SPC chart, first-order digital filter).

FDC: Fault detection and control.
Feedback control: Uses measurements about the current process results to decide how to change the process for the next sampling period. Thus, feedback control drives the average value to target (i.e., drives C_{pk} to equal C_p). Used to compensate for expected disturbances.
Feedforward: Control uses measurements on incoming materials, process, or equipment to decide how to change the process for the current process. Feedforward control, because it accounts for incoming variations, can improve the C_p value, i.e., it turns apparently random variation into non-random variation, which can be compensated for. Used to compensate for measured disturbances that have been modeled.
Gain: The change in the output for a unit change in the input.
IID$N(\mu,\sigma)$: Identically independently distributed normal means all values belong to the same distribution which is a normal (Gaussian) distribution with mean of μ and standard deviation of σ.
Look-aheads: Wafers from a production lot processed and analyzed before the rest of the production lot is processed.
MBPC: Model-based process control.
WIP: Work in progress.
MES: Manufacturing execution system (used for WIP management, material flow control).
Metrology: Measurement science; the act of measuring; the measurement process.
Pilot: Non-sellable wafer, may or may not have topography.
Power: The probability to detect a particular size of change ($=1-\beta$). Term encountered in sample size calculations in statistical tests and SPC.
Process model: A model that predicts the inherent response of the process to input changes.
SPC: Statistical process control.
SRC: Semiconductor research company (research consortia that funds universities).
Trace: Real-time signal over time. Usually implies signals from processing equipment, but source could also be add-on sensors.
Type I error: False positive; to detect an abnormality when one has not occurred.
Type II error: False negative; to not detect an abnormality when one has occurred.
WECO: Supplementary run rules applied in conjunction with Shewart SPC. These rules are generally known as the Western Electric (WECO) rules in recognition of the source of their well-known applications. These rules simultaneously decrease type II error and increase type I error.
PCC: Predictor corrector controller. Filter for continuous drifts.
Pilot: Non-sellable wafer. Used when processing or metrology could cause out of specification results, produce contamination, or create defects. Also used for machine recovery and conditioning when the equipment must be operated, since most plasma and deposition machines do not allow non-wafer processing.
Qual: Short for qualification. Also, a run, usually using pilot wafers, used to qualify a machine to run a particular process(es) after maintenance or when switching between processes.
Qual plan: A methodology to qualify a process for use in the manufacture of production material.
Random yield loss: Yield loss due to contamination, particles (as opposed to systematic yield loss)
Run: A single processing on a piece of equipment or the processing of all the wafers in the lot, which may consist of several processings on single-wafer equipment.
Systematic yield loss: Unintentional and intentional misprocessing (as opposed random yield loss).

References

1. Butler, S. W., R. York, M. H. Bennett, and T. Winter. "Semiconductor Factory Control and Optimization." *Wiley Encyclopedia of Electrical and Electronics Engineering*, Vol. 19, 59–86. New York: Wiley, 1999.
2. Two-Year Keithley Study Eyes Process Control, *WaferNews*, Vol. 4.29, 3, 6. 28 July 1977.

3. Box, G., and A. Luceno. *Statistical Control by Monitoring and Feedback Adjustment.* Wiley Series in Probability and Statistics, Wiley, 1997. Also, course from University of Wisconsin–Madison College of Engineering, Feedback Adjustment for SPC, How to Maximize Process Capability Using Feedback Adjustment, Box, G., J. Hunter, and S. Bisgaard.
4. Seborg, D. E., T. F. Edgar, and D. A. Mellichamp. *Process Dynamics and Control*. New York: Wiley, 1989.
5. Vander Wiel, S. A., W. T. Tucker, F. W. Faltin, and N. Doganaksoy. "Algorithmic Statistical Process Control: Concepts and an Application." *Technometrics,* 34, no. 3 (1992): 286–97.
6. Tucker, W. T., and F. W. Faltin. "Algorithmic Statistical Process Control: An Elaboration." *Technometrics* 35, no. 4 (1993): 363–75.
7. MacGregor, J. F. "Interfaces between Process Control and On-Line Statistical Process Control." *AIChE Comput. Syst. Technol. Div. Commun.* 10, no. 2 (1987): 9–20.
8. Box, G. and T. Kramer. "Statistical Process Monitoring and Feedback Adjustment: A Discussion." *Technometrics,* 34, no. 3 (1992): 251–67.
9. Hoerl, R. W. and A. C. Palm. "Discussion: Integrating SPC and APC." *Technometrics* 34, no. 3 (1992): 268–72.
10. MacGregor, J. F. "Discussion." *Technometrics,* 34, no. 3 (1992): 273–5.
11. Tucker, W. T. "Discussion." *Technometrics,* 34, no. 3 (1992): 275–7.
12. Vander Wile, S. A., and S. B. Vardeman. "Discussion." *Technometrics,* 34, no. 3 (1992): 278–81.
13. Wardrop, D. M., and C. E. Garcia. "Discussion." *Technometrics,* 34, no. 3 (1992): 281–2.
14. Box, G., and T. Kramer. "Response." *Technometrics,* 34, no. 3 (1992): 282–5.
15. Muthukrishnan, S., and J. Stefani. "SCFab Model-Based Process Control Methodology: Development and Deployment for Manufacturing Excellence." *TI Tech. J.* 13, no. 5 (1996): 9–16.
16. Butler, S. W., J. Stefani, and G. Barna. "Application of Predictor Corrector Controller to Polysilicon Gate Etching." In *Proceedings of the American Control Conference*, 3003. Piscataway, NJ: IEEE, 1993.
17. Butler, S. W. "Process Control in Semiconductor Manufacturing." *J. Vac. Sci. Tech. B* 13, no. 4 (1995): 1917–23.
18. Butler, S. W., J. Hosch, A. Diebold, and B. Van Eck. "Sensor Based Process and Tool Control." *Future Fab Int.* 1, no. 2 (1997): 315–21.
19. Butler, S. W., and T. F. Edgar. "Case Studies in Equipment Modeling and Control in the Microelectronics Industry." In *Proceedings of the Fifth Conference on Chemical Process Control (CPC V) Assessment and New Directions for Research. AIChe Symposium Series No. 316*, Vol. 93, 133–44. CACHE, American Institute of Chemical Engineers, 1997.
20. Semiconductor Industry Association (SIA). *The National Technology Roadmap for Semiconductors.* San Jose, CA: SIA, 1997, (408)246-2830, http://www.semichips.org
21. Kraft, C. "TI's Statistic's Needs." *SEMATECH Statistics Workshop.* Austin, Nov 1988.
22. Czitrom, A. V., and K. Horrell. "SEMATECH Qual Plan: A Qualification Plan for Process and Equipment Characterization." *Future Fab Int.* 1, no. 1 (1996): 45.
23. Clark, W., K. Horrell, T. Rogelstad, and P. Spagon. *SEMATECH Qualification Plan Guidelines for Engineering.* SEMATECH DOC ID No.: 92061182B-GEN, SEMATECH, 1995; Horrell. K. *SEMATECH Qualification Plan Overview.* SEMATECH DOC ID No.: 91050538B-GEN, SEMATECH, 1993.
24. See http://www.domainmfg.com/ for information on Starfire by Domain Manufacturing Corp., acquired by Brooks Automation 6/99.
25. Mason, R. L., R. F. Gunst, and J. L. Hess. *Statistical Design and Analysis of Experiments with Applications to Engineering and Science.* 2nd ed. New York: Wiley-InterScience, 2003.
26. Naugib, H. "The Implementation of Total Quality Management (TQM) in a Semiconductor Manufacturing Operation." *IEEE Trans. Semicond. Manuf.* 6, no. 2 (1993): 156.
27. Grant, E. L., and R. S. Leavenworth. *Statistical Quality Control*. New York: McGraw-Hill, 1988.
28. Drain, D. *Statistical Methods for Industrial Process Control*. New York: Chapman and Hall, 1997.
29. Shewhart, W. A. *Economic Control of Quality of Manufactured Product*. New York: Van Nostrand, 1931.

30. Vicker, K. "Advanced Process Control in the Fab." *SEMATECH Advanced Equipment and Process Control Workshop*, 338–51. Lake Tahoe, CA: SEMATECH, 1997.
31. Guldi, R. L., C. D. Jenkins, G. M. Damminga, T. A. Baum, and T. A. Foster. "Process Optimization Tweaking Tool (POTT) and Its Applications in Controlling Oxidation Thickness." *IEEE Trans. Semicond. Manuf.* 2 (1989): 54–9.
32. Wheeler, D. J., and D. S. Chambers. *Understanding Statistical Process Control,* 2nd Ed. Knoxville, TN: SPC Press, 1992.
33. Mandel, J. "The Regression Control Chart." *J. Qual. Technol.* 1, no. 1 (1969): 1–9.
34. Sachs, E., A. Hu, and A. Ingolfsson. "Run by Run Process Control: Combining SPC and Feedback Control." *IEEE Trans. Semicond. Manuf.* 8, no. 1 (1995): 26–43: Boning, D., W. Moyne, T. Smith, J. Moyne, and A. Hurwitz. "Practical Issues in Run by Run Process Control." In *Proceedings of IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 201–8, 1995.
35. Mozumder, P. K., and G. G. Barna. "Statistical Feedback Control of a Plasma Etch Process." *IEEE Trans. Semicond. Manuf.* 7, no. 1 (1994): 1–11.
36. Ling, Z.-M., S. Leang, and C. J. Spanos. *In-Line Supervisory Control in a Photolithographic Workcell.* SRC publication C91008; also SPIE, 921 (1988):258: Bombay, B. J., and C. J. Spanos. "Application of Adaptive Equipment Models to a Photolithographic Process." *SPIE Technical Symposium on Microelectronic Processing Integration*, September 1991: Leang, S., S-Y. Ma, J. Thompson, B. J. Bombay, and C. J. Spanos. "A Control System for Photolithographic Sequences." *IEEE Trans. Semicond. Manuf.* 9, no. 2 (1996): 191–207.
37. Toprac, A. *Run-to-Run Control of Poly-Gate Etch SEMATECH Advanced Equipment and Process Control Workshop*, 434–40. Lake Tahoe, CA: SEMATECH, 1997.
38. Gerold, D. "Run-to-Run Control Benefits to Photolithography." *SEMATECH Advanced Equipment and Process Control Workshop*. 104–12. Lake Tahoe, CA: SEMATECH, 1997, suppl.
39. Stefani, J. A. "Practical Issues in the Deployment of a Run-to-Run Control System in a Semiconductor Manufacturing Facility." *The 1999 SPIE International Symposium on Microelectronic Manufacturing Technologies*. Edinburgh, Scotland, May 1999.
40. Hunter, J. S. "The Exponentially Weighted Moving Average." *J. Qual. Technol.* 18 (1986): 203–10.
41. Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control.* Englewood Clifs, NJ: Prentice Hall, 1994.
42. Butler, S. W., and J. Stefani. "Supervisory Run-by-Run Control of Polysilicon Gate Etch Using In Situ Ellipsometry." *IEEE Trans. Semicond. Manuf.* 7, no. 2 (1994): 193.
43. Butler, S. W., J. Stefani, M. Sullivan, S. Maung, G. Barna, and S. Henck. "An Intelligent Model Based Control System Employing In Situ Ellipsometry." *J. Vac. Sci. Tech. A,* 12, no. 4 (1994): 1984–91.
44. Marlin, T. E., and A. N. Hrymak. "Real-Time Operations Optimization of Continuous Processes." In *Proceedings of the Fifth Conference on Chemical Process Control (CPC V) Assessment and New Directions for Research. AIChE Symposium Series No. 316*, Vol. 93, 156–64. CACHE, American Institute of Chemical Engineers, 1997.
45. Harris, T. J., and J. F. MacGregor. "Design of Multivariate Linear-Quadratic Controllers Using Transfer Functions." *AIChE J.* 33 (1987): 1481–95.
46. Smith, T., D. Boning, J. Stefani, and S. W. Butler. "Run by Run Advanced Process Control of Metal Sputter Deposition." *IEEE Trans. Semicond. Manuf.* 11, no. 2 (1998): 276.
47. Bohn, R. "The Impact of Process Noise on VLSI Process Improvement." *IEEE Trans. Semicond. Manuf.* 8, no. 3 (1995): 228–38.

Further Reading

Conferences and Supporting Organizations

Integrated Metrology Consortia, http://www.integratedmeasurement.com
Electrochemical Society, Inc., http://www.electrochem.org
American Vacuum Society (AVS) Manufacturing Science and Technology Group (MSTG), http://www.cems.umn.edu/~weaver/mstg/mstgsubway.html
International Symposium on Semiconductor Manufacturing (ISSM), http://www.issm.com
Advanced Semiconductor Manufacturing Conference (ASMC), http://www.semi.org/Education/asmc/main.html
SPIE Microelectronic Manufacturing, http://www.spie.org/info/mm

SC Control and Control Software

University of Michigan Controls Group, http://www.engin.umich.edu/research/controls
Berkeley Computer Aided Manufacturing (BCAM), http://bcam.berkeley.edu
Maryland University The Institute for Systems Research, http://www.isr.umd.edu
SEMATECH and MIT Run by Run Benchmarking, http://www-mtl.mit.edu/rbrBench
Triant Technologies Inc., http://www.triant.com
Semy, http://www.semyh.com
ObjectSpace, http://www.objectspace.com
Domain Manufacturing Corp. (formerly, BBN Domain Corp; acquired by Brooks Automation 6/99), Cambridge, MA, http://www.domainmfg.com
Umetrics, Winchester, MA, http://www.umetri.se (also good Chemometrics links).
Brookside Software, http://www.brooksidesoftware.com
Brooks Automation, Richmond, BC, Canada, http://www.brooks.com/bac.htm
Fastech (Acquired by Brooks Automation 9/98), http://www.fastech.com
Real Time Performance, Sunnyvale, CA, http://www.rp.com (no longer in business, but has code).
Adventa (ControlWORKS, ProcessWORKS, WORKS), Dallas, TX, http://www.adventaCT.com
Voyan Technology, Santa Clara, CA.
PRI Automation, Inc., Billerica, MA, http://www.pria.com
Bakshi, V. Fault Detection and Classification (FDC) Software Benchmarking Results, SEMATECH Technology Report 97123433A-TR, 1998.
Bakshi, V. Fault Detection and Classification Software for Plasma Etchers: Summary of Commercial Product Information. SEMATECH Technology Report 97083337A-XFR, 1997.

Manufacturing Execution Systems (MES)/Computer Integrated Manufacturing (CIM)/Equipment Integration Automation: Software Used to Run and Track Fab, Perform SPC, etc.

Fastech (Acquired by Brooks Automation 9/98), http://www.fastech.com
Real Time Performance, Sunnyvale, CA, http://www.rp.com (no longer in business, but has code).
Consillium (Acquired by Applied Materials), http://www.consilium.com
Promis (Acquired by PRI Automation), http://www.promis.com

Byrd, T. and A. Maggi. "Challenges to Plug and Play CIM." *Future Fab Int.*: 77–81.
Greig, M. and A. Weber. "AMD & ObjectSpace, Inc." *Future Fab Int.*: 73–74.
Data Analysis, Data Warehousing, Data Mining, Bit Mapping, Wafer Tracking, etc.
Knight's Technology, Sunnyvale, CA, http://www.knights.com
DYM, Bedford, MA, http://www.dym.com
LPA Software, South Burlington, VT, (802) 862-2068.
Quadrillion, Company Information, http://www.quadrillion.com/quadinfo/htm
Device Ware Corporation, http://www.dware.com
Maestro, Data Management [JJT Inc.], http://www.jjt.com/data.man.html
Sleuthworks, http://www.sleuthworks.com/doc
SAS, http://www.sas.com
NIST and SEMATECH have created an Engineering Statistics Internet (ESI) Handbook. Check the SEMATECH web page or contact Chelli, Zey@SEMATECH.Org

General Semiconductor References That May Contain Control References on Occasion

SEMATECH, www.sematech.org
I300I (dedicated to 300 mm issues), www.i300i.org
National Technology Roadmap, http://www.sematech.org/public/roadmap/index.htm
Semiconductor Subway, http://www-mtl.mit.edu/semisubway.html
Semiconductor Equipment and Materials International (SEMI), http://www.semi.org
Semiconductor Research Corporation (SRC), http://www.semi.org
Semiconductor International, http://www.semiconductor-intl.com
Solid State Technology, http://www.solid-state.com
Semiconductor Online, http://www.semiconductoronline.com
Semiconductor SuperSite.Net, http://supersite.net/semin2/docs/home.htm
FabTech, www.fabtech.org
TechWeb, http://www.techweb.com
Semiconductor Process Equipment and Materials Network, http://www.smartlink.net/~bmcd/semi/cat.html
semiconductor.net–The semiconductor manufacturing industry resource for products, services and information, http://www.semiconductor.net
SemiSource, Semiconductor Resource Guide, published annually by Semiconductor International.
Solid State Technology Resource Guide, published annually by Solid State Technology.
American Vacuum Society (AVS) Buyers Guide, http://www.aip.org/avsguide

Additional Compensation Control and Controller Monitoring Articles

C Kraft. U.S. patent 5,528,510, Equipment Performance Apparatus and Method, issued 6.19.96.
Harris, T. J. "Assessment of Control Loop Performance." *Can. J. Chem. Eng.* 67 (1989): 856–61.
Scher, G. M. "Wafer Tracking Comes of Age." *Semicond. Int.* (1991): 126–31.
Stefani, J. A. "Practical Issues in the Deployment of a Run-to-Run Control System in a Semiconductor Manufacturing Facility." *The 1999 International Symposium on Microelectronic Manufacturing Technologies*. Edinburgh, Scotland, May 1999.

24

In-Line Metrology

24.1 Introduction **24**-1
Measurement Precision to Process Tolerance Ratio vs. Resolution • Manufacturing Sensitivity Analysis

24.2 Metrology for Lithography Processes: Critical Dimension Measurement and Overlay Control **24**-5
Critical Dimension Measurement and Calibration • Overlay Process Control

24.3 Metrology for Front End Processes **24**-20
Ellipsometric Measurement of Gate Dielectric Film Thickness • Electrical Measurement of Gate Oxide Thickness • New Methods of Measuring Gate Dielectric Thickness and Nitrogen Concentration • Doping Process Control • Metrology for Measurement of Stress Enhanced Carrier Mobility

24.4 Interconnect Process Control **24**-36
Interconnect Film Thickness • Ex-Situ Chemical Mechanical Polishing Process Control–Film Flatness and Quality

24.5 In-FAB FIB **24**-49

Acknowledgments **24**-49

References **24**-50

Alain C. Diebold
SEMATECH

Abstract

This chapter on ex-situ metrology covers the in-line and at-line measurements used to control processes in pilot line fabrication or in volume manufacturing of silicon based integrated circuits. Metrology for front end (transistor) processes, lithography, and on-chip interconnect fabrication technologies are all described. As a basis for further discussion, measurement precision and resolution are described, and the measurement precision to process tolerance ratio is used to evaluate metrology capability for statistical process control.

24.1 Introduction

Metrology is an integral part of the development and manufacture of integrated circuits. The International Technology Roadmap for Semiconductors (ITRS) defines metrology as including both the off-line materials characterization and in-line measurement technologies [1]. In-line metrology is discussed in this chapter, while other chapters are devoted to other aspects of metrology. Other chapters of interest include those covering process control, off-line materials characterization, in-situ sensor based metrology, and defect detection and control technology.

In-line metrology covers measurement needs for manufacturing control of transistor and on-chip interconnect fabrication including lithography. A high level view of the frequency of metrology

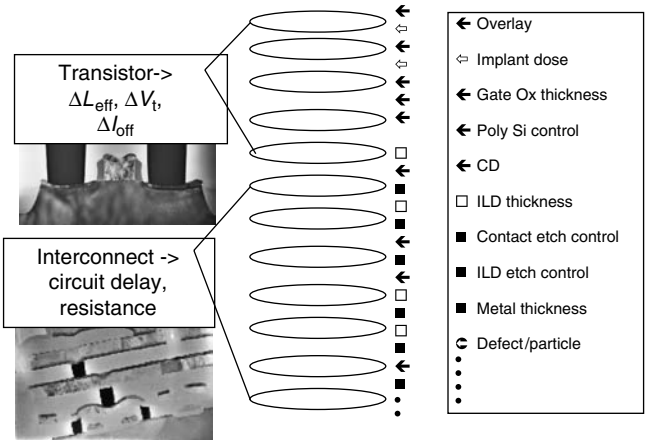


FIGURE 24.1 Overview of the applications of metrology during integrated chip manufacturing.

measurements during manufacture is shown in Figure 24.1. First, the concepts of measurement precision and resolution are reviewed in Section 24.1. Section 24.2 covers critical dimension (CD) and overlay. Section 24.3 discusses gate dielectric thickness measurement and doping process control. Interconnect metrology is covered in Section 24.4, and other measurement tools such as focused ion beam (FIB) and scanning electron microscopy are briefly covered in Section 24.5.

24.1.1 Measurement Precision to Process Tolerance Ratio vs. Resolution

The measurement precision to process tolerance (P/T) ratio is an accepted metric for evaluating the ability of an automated metrology tool to provide data for statistical process control (SPC) [2]. While there is no generally accepted estimator for precision [3] an ideal measure for use in process control would be a function of the total measurement system error variability, σ_M^2. A methodology for estimating P/T in this fashion can be found in SEMI E89, Guide for Measurement System Analysis [4]. Using this or a similar approach, major sources of measurement variability are enumerated and a designed experiment is employed to quantify the variance components associated with each source. σ_M^2 is a combination of the short term and the long term sources of variation. Repeatability, σ_r represents a lower limit to σ_M^2. It is estimated by the sample standard deviation of repeated measurements made under identical conditions, or in the case of a designed experiment, as the square root of the mean squared error [3,4]. Reproducibility, σ_R, is the variation that results when measurements are made under different conditions such as reloading the wafer on a different day [3,4]. It may include multiple sources of variability and is expressed as the square root of the sum of the variance components for all sources of measurement system variation. Note: SEMI E89 treats repeatability as part of reproducibility, making $\sigma_M^2 = \sigma_R^2$. This approach is not universal. Some treat repeatability and reproducibility as mutually exclusive, in which case $\sigma_M^2 = \sigma_R^2 + \sigma_r^2$.

Precision is calculated as $C\sigma_M$, where $C=6$ is a common choice for processes with both upper and lower limits and $C=3$ for processes with one-sided limits. Contamination limits represent a process with a one-sided (upper) process limit. Process tolerance is the range of allowed values of the process variable: upper process limit − lower process limit (UL − LL) for two-sided processes and (UL − T) or (T − LL), where T is a target value, for one-sided processes. The measurement precision, σ, to P/T ratio is either: $6\sigma/(\text{UL}-\text{LL})$, $3\sigma/(\text{UL}-T)$, $6\sigma/(T-\text{LL})$ [4]. Although P/T should be less than 10%, a value of 30% is often allowed.

Unfortunately, the measurement precision used to determine P/T is often an approximation of the true precision. Determination of the true P/T ratio for the process range of interest requires careful

implementation of the *P*/*T* methodology. This means that many of the measurement conditions should be varied to determine measurement stability before setting final measurement conditions when determining precision [4]. Varying time between repeated measurement allows one to observe short term issues with repeatability. In addition, reference materials should have identical feature size, shape, and composition to the processed wafer being measured. Often, there are no reference materials that meet this requirement, and *P*/*T* ratio and measurement accuracy are determined using best available reference materials. One key example is the lack of an oxide thickness standard for sub 5 nm SiO_2 and nitrided oxides.

When the reference material has significantly different properties (e.g., thickness), then the precision may not be representative of the precision associated with the product wafer measurement due to non-linearity. Again, the example of CD or film thickness measurements is useful. The precision associated with measurement of a sub 2 nm gate oxide may be different than that associated with a 10 nm oxide film. If the true precision is large enough, it could mean that the metrology tool has insufficient resolution to distinguish changes over the process range of interest. One way to assure that the metrology tool has adequate resolution to determine the true *P*/*T* capability by using a series of standardized, accurate reference materials over the measurement range specified by the ULs and LLs. In Figure 24.2, we depict the difference between precision and bias. In Figure 24.3, we show how the multiple reference materials approach might work. This approach is not used. Instead, in-line metrology uses a reference wafer that is fabricated using the typical process flow. These "golden" wafers have exact materials and dimensions from the process step of interest. Often, only one suitable reference material is used for *P*/*T* determination.

The measurement of the thickness of the transistor gate dielectric at the 45 nm technology generation is expected to be difficult in the manufacturing environment. By the 45 nm technology generation, silicon oxynitride will be replaced by a higher dielectric constant material. Dielectric thickness will be written in terms of the thickness that the layer would have if it were silicon dioxide. This is known as the equivalent oxide thickness (EOT). For single layer films, the EOT can be calculated by multiplying the physical film thickness by the ratio of the dielectric constant of silicon dioxide ($\kappa(SiO_2) = 3.9$) to that of the high κ material. For high κ films with an interfacial layer of silicon dioxide of thickness t_{int}, the EOT = $t_{int} + (3.9/\kappa)t_{high}\kappa$. If the gate dielectric EOT is 0.7 nm thick and the process tolerance is 4% for 3σ (process variation), then $P/T = 10\% = 6\sigma/(0.1\text{ nm})$ which gives a measurement variation $3\sigma =$ 0.0028 nm. The size of an atomic step on silicon is $\sim$0.15 nm and atomically flat terraces on specially prepared Si (001) are about 100–200 nm wide. The width of terraces after typical processing such as sacrificial oxidation and gate oxide growth is unknown. Thus, some have pointed to the issue of measuring film thickness to less than an atomic width. This is not an issue because the measurement can be considered to be the determination of the average thickness of a perfectly flat layer. This type of measurement precision requires analysis of a large area that averages over atomic steps at the interface

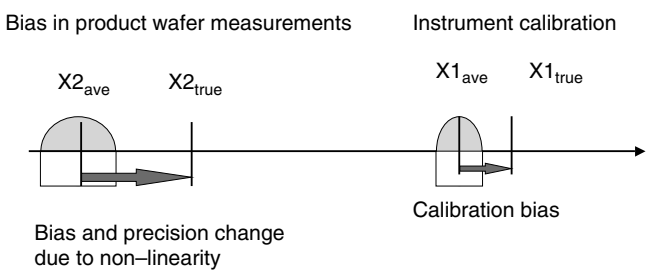


FIGURE 24.2 Measurement non-linearity measurement non-linearities can result in bias (difference between true and measured value) changes between the calibrated value and the range of interest. It is also possible that the bias can change inside the measurement range of interest.

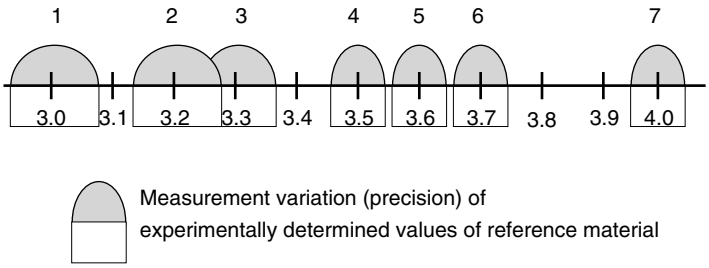


FIGURE 24.3 Demonstration of resolution based on precision associated with measurement of a series of reference materials over the process specification range for this example let us assume that the process tolerance (also called process specifications) is from 3.0 to 4.0 nm. The measurement precision at 3σ (variation) is shown for reference materials inside the process range. The experimental P/T capability observed using reference materials 4, 5, and 6 indicates that a single measurement of a 3.6 nm is different from one at 3.5 or 3.7 nm. Thus this fictitious metrology tool can resolve those values. According to the precision shown for reference materials 2 and 3, the tool is not able to resolve 3.2 nm from 3.3 nm at 3σ.

and other atomic variations. Local variations in thickness of a 2 nm film will be a much larger percentage of total thickness than it would be for a 10 nm film. Therefore, reproducibility of stage location affects the precision of a thickness measurement especially for a small sampling area around 1 μm in diameter. Some metrologists have called for smaller area (<1 μm) measurements. The need for high precision makes this very difficult and high precision with small spot instruments (0.9 μm) are achieved by averaging over a larger area. Even electrical measurements done using capacitors average over larger areas.

For contamination measurements and a $P/T = 30\%$, $\sigma = (\text{UL} - \text{L})/10$. Achieving a measurement variation of $(\text{UL} - \text{L})/10$ requires that the detection limit must be at or below $(\text{UL} - \text{L})/10$ [3]. Chemists have their own definitions of resolution, detection limit, and quantifiable detection limit which must all be considered [5]. For convenience, these topics are illustrated by discussing off-line chemical analysis of trace contaminants in a liquid. Typically, the detection limits for trace amounts of contamination in a liquid vary due to changes in measurement sensitivity. Detection limits are ultimate values for optimum circumstances. When the liquid is analyzed by inductively coupled plasma mass spectrometry (ICP-MS), the detection of one element can be interfered with by the presence of another element. The quantifiable detection limit is the limit at which reliable, repeatable detection is possible [5]. Resolution requirements are not well defined, and thus experience again provides some guidance.

24.1.2 Manufacturing Sensitivity Analysis

Ultimately, the electrical performance of the integrated circuit results from the electrical properties of the transistors and the interconnect structures. The most extensive modeling of how variations in physical properties affect electrical performance has been done for transistors [6]. The effect of small changes in key physical parameters such as gate length, gate dielectric thickness, and doping dose on key electrical parameters such as leakage current and threshold voltage were modeled for a 180 nm transistor [6,7]. Although similar modeling has been done for subsequent generations of transistors, most results are not yet published. The goal was to determine the impact of an increase in the range of a variable on the range of the electrical parameters. Modeling for this purpose is known as manufacturing sensitivity analysis. This type of information is often helpful in prioritizing metrology and understanding the electrical impact of process variation.

Another type of model-based sensitivity analysis has been used to relate the electrical test signature to physical defect type for interconnect structures [8,9]. This is sometimes called defect to fault mapping.

It is usually associated with off-line materials characterization and failure analysis laboratories that employ FIB systems, and thus it is not discussed at all in this chapter.

24.2 Metrology for Lithography Processes: Critical Dimension Measurement and Overlay Control

Critical dimension and overlay measurements control two of the most important parts of manufacturing: feature delineation and the stacking of layers. In this section, CD and overlay measurement technology and application are reviewed. Detailed chapters on each CD measurement method and on overlay measurement can be found in the Handbook of Silicon Semiconductor Metrology [10].

Control of CD including line edge roughness (LER) across the die and across the wafer starts with control of the range of CDs across the photomask. Since photomasks have glass substrates, they easily charge during scanning electron microscopy based CD measurement. Thus, CD measurements on photomasks requires special considerations which are discussed below. CD metrology is also done on the patterned photoresist structures as well as after the features are etched. Issues associated with these measurement steps are also covered.

24.2.1 Critical Dimension Measurement and Calibration

Today, both critical dimension scanning electron microscopes (CD-SEM) and scatterometry are used during volume manufacture to control the process range for line width and contact via diameter (area). Cross-sectional SEM images and CD-SEM are used to evaluate process conditions after a change in recipe. Electrical measurement of the effective transistor gate length and metal line width are also done using test structures. Electrical measurements support tighter control of electrical transistor parameters that is possible with current physical CD-SEM measurements. Lithography process control using CD-SEM or scatterometry measurements can be done after exposure and development of the resist, thus permitting reprocessing of wafers that have CD values outside of process tolerance limits. CD is also measured after etch of the poly-silicon transistor gate or the trenches and vias used for interconnect. In this section, we will discuss CD-SEM and scatterometry tool and measurement issues, calibration, and electrical CD measurement.

24.2.1.1 CD-SEM Tools and Measurement Issues

A CD-SEM is a scanning electron microscope that has been designed for low voltage ($< \sim 1$ keV) measurement of line width. In Figure 24.4, we illustrate the operation of a CD-SEM. In Figure 24.5, show a typical CD-SEM design. Conventional wisdom leads one to believe that low voltage operation is the only way to prevent damage to the integrated circuit (IC) chip, and at this time, commercial CD-SEMs operate at voltages between roughly 1 keV and a few 100 eV. Although CD-SEMs differ from in-line and laboratory SEMs in that very short working distances (distance between wafer and SEM electron beam lens) are used to optimize rapid, low voltage imaging, new lens designs are reducing the differences between systems [11]. In order to maintain optimum and reproducible fields over the sample, sample tilt is not used in CD measurement. Recently introduced CD-SEMs have become more capable of measuring sidewall angle. Using the electron lens to tilt the electron beam, sidewall shape and angle can be measured as discussed below. Multiple angles allow determination of sidewall angle through geometric considerations. This increases the rate at which the sample stage can scan the wafer. Precision sample stages that can locate specific wafer coordinates combined with image pattern recognition capability allow automated CD measurement at specific locations on selected die across the wafer. Two components of the SEM hardware are considered key to higher resolution and greater CD precision. First, the final electron beam lens has been designed so the magnetic field extends beyond the lens and around the part of the sample being imaged. This type of lens is referred to as an external extended field lens or snorkel [12]. The second development is the advanced secondary electron detector. The extended field also collects the secondary

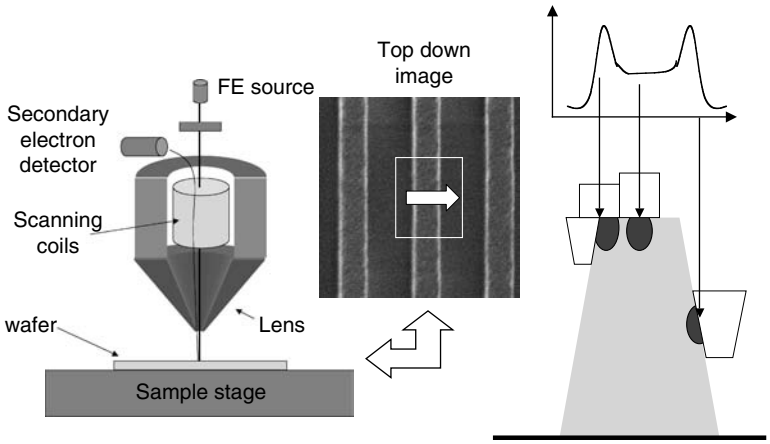


FIGURE 24.4 How a critical dimension scanning electron microscope (CD-SEM) works. A line scan is averaged over a length of the measured line using a specially designed SEM. The increase in secondary electron intensity at the line edge is shown. The emitted electron regions are highlighted in green. Emitted electron intensity is the largest at the edge of a line (point 1) and the least at a flat surface (point 2). Sharp features such as the line edge have more surface area for electron emission. More emitted electrons (point 3) will come from sidewalls than from flat areas (point 2).

electrons which are then passed through a Wein filter that results in energy dispersion prior to detection [13]. This allows the ultra rapid collection at 5× to 10× television (TV) scanning rates.

A discussion of the origin of the CD-SEM signal and its interpretation as a line width will assist understanding CD-SEM measurements. A recent review of SEM based CD measurement, calibration, and SEM matching is a recommended Ref. [13]. A line trace of detected electron intensity vs. sample position, a SEM image of a line structure and a SEM cross-section of a line structure are shown in Figure 24.4. Both low energy secondary electrons and backscattered (inelastically) scattered electrons can be used separately or in combination to produce an image. Secondary electrons are low energy electrons emitted from the valence band of the sample after excitation by the primary electron beam [14], and they are often labeled as SE-1 electrons [13,14]. The SE-1 yield is a function of the sample material, sample shape, and the energy of the primary electron beam. Sharp features and side walls will be able to emit more secondary electrons than a flat surface. The secondary electron coefficient is a measure of the number of secondary electrons per incident electron from the primary beam. For most materials, the secondary electron coefficient peaks between 1 and 2 keV electron beam energies [13]. Electrons that scatter off sample atoms and out of the sample are known as backscattered electrons [13]. Some of these electrons will have lost energy through typically several collisions inside the sample and are thus inelastically scattered. Some authors [13,14] refer to these electrons as SE-3 secondary electrons, and others seem to prefer not designating these as secondary but as backscattered electrons [13,14]. Other backscattered electrons, known as elastically scattered electrons, hit atoms close to the surface and escape before any energy is lost. These electrons are referred to as SE-2 by some authors [13]. There will be more backscattered electrons from small sharp features and sidewalls than flat surfaces. The emitted electron intensity will be higher at the edge of the line structure as shown in Figure 24.4. In older CD-SEMs, the detected signal is a combination of both the low energy secondary electrons (SE-1) and the backscattered electrons (SE-2 and SE-3). The way in which these contributions are mixed will determine the shape of the signal and how well the line edge algorithm is able to measure line edge. The signal shown in Figure 24.4 demonstrates that one cannot interpret the signal intensity in terms of the sample height. The peak intensity cannot be interpreted as the line edge position [14,15].

Critical dimension measurement is done by averaging over a large number of line scans across a length of line deemed representative. (The SEM image is composed of many line scans with the vertical

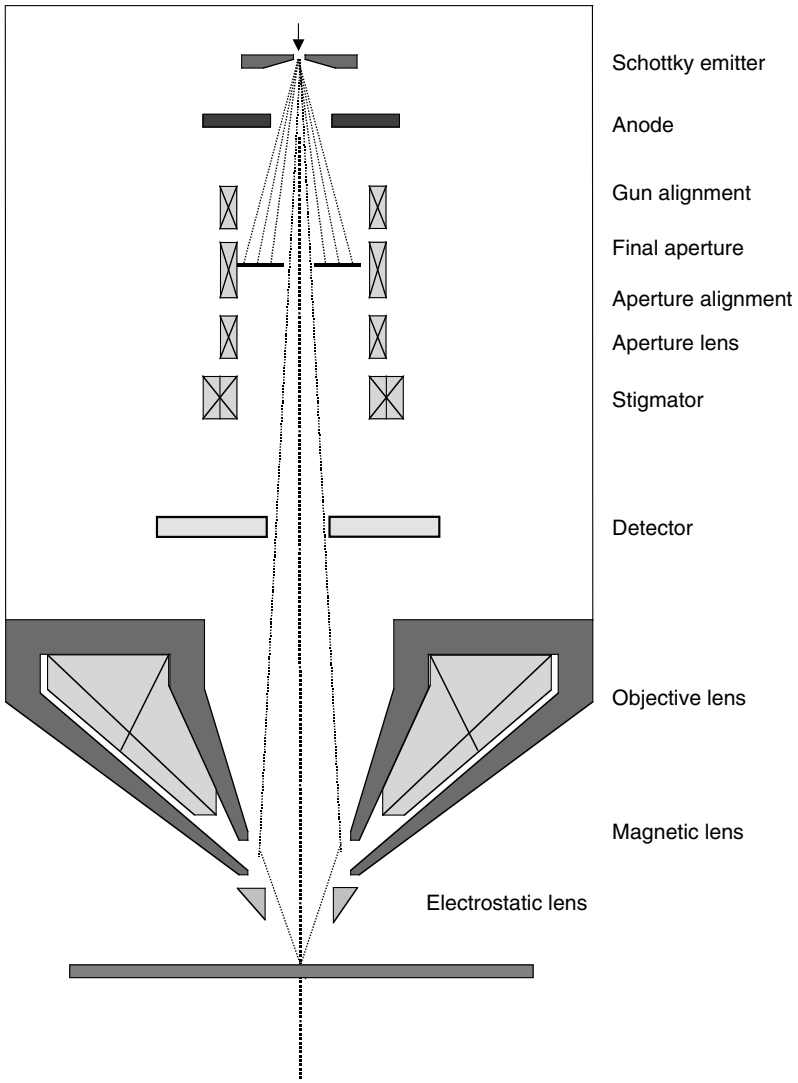


FIGURE 24.5 Diagram of CD-SEM lens. Applied materials CD-SEM uses a combined electrostatic/magnetic lens and has a through the column emitted electron detector. Figure courtesy Bob Burkhardt of applied materials.

(or horizontal) position of each scan moved by a known increment.) This averages out some of the edge roughness effects and sample variation. At the time of writing this chapter, many different methods of determining edge position exist, and each supplier can chose a method that optimizes the precision of their own instrument. Some of the algorithms are shown in Figure 24.6. It has been shown that the linewidth can vary by 100 nm according to the algorithm selected [12]. Clearly, CD-SEM measurements must be calibrated by a method that is more independent of sample shape and line materials.

The above paragraph points to the need for a standardized algorithm so that measurement equipment can be matched. This is critical for manufacturing metrology. The above discussion also points to the need for a fundamental model that relates the true feature shape to the signal inside the CD-SEM [15]. Through Monte Carlo modeling, the secondary electron intensity of a linescan can be related to lineshape. In addition, the effect of the width of the electron beam can be removed. This modeling has also been used to improve the CD precision (nist paper and Schlumberger paper). The use of very low

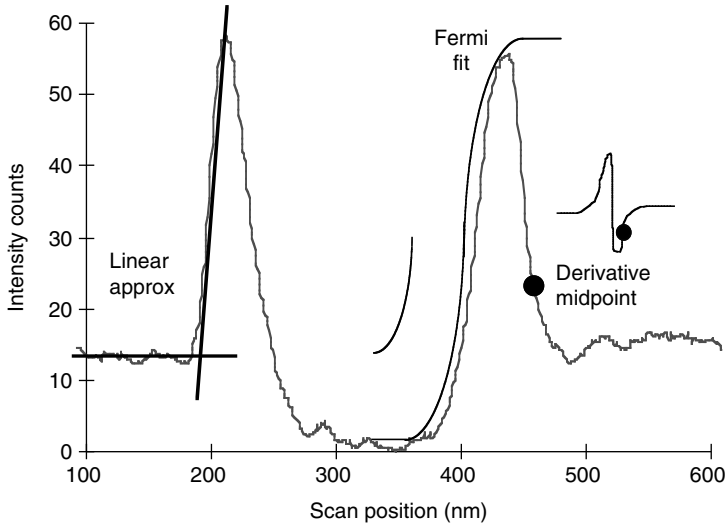


FIGURE 24.6 The algorithm used for determining the line edge impacts the measured value of line width. The red line is the signal from the SEM line scan. Three algorithms are shown: linear approximation, Fermi Fit, and derivative method. Figure 24.6 first appeared in reference 12, and it was used with the authors' permission.

voltages also changes linescan shape as shown in Figure 24.7. Once again, the algorithm that extracts linewidth from linescan information must be altered to allow for changes in linescan shape. Over the past several years, printed linewidth has fallen below 50 nm. These resist features often have rounded tops, and a line scan over this type of feature does not have the same shape as larger rectangular lines. This is shown in Figure 24.8. In order to determine linewidth, new algorithms were developed. These algorithms use the same sort of edge determination approach as that used for measurement of beam width using a sharp edge feature [16].

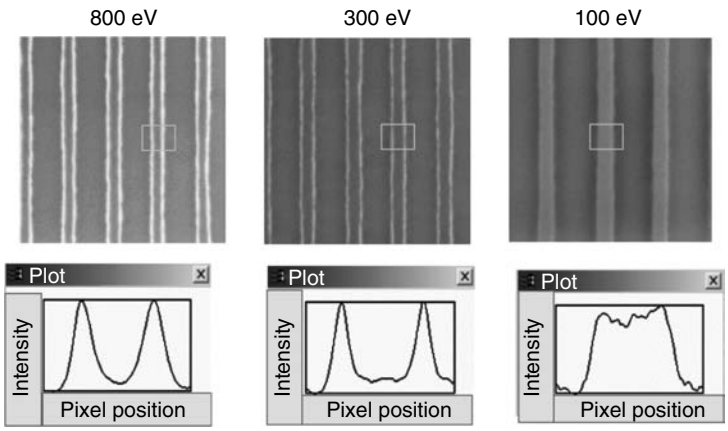


FIGURE 24.7 Line scan shape change during ultra low voltage CD-SEM. Decreasing voltage has a significant change in the linescan intensity vs. position. The high intensity "lobes" normally observed at the line edge disappear at ultra-low voltage. Figure courtesy Neal Sullivan (Schlumberger).

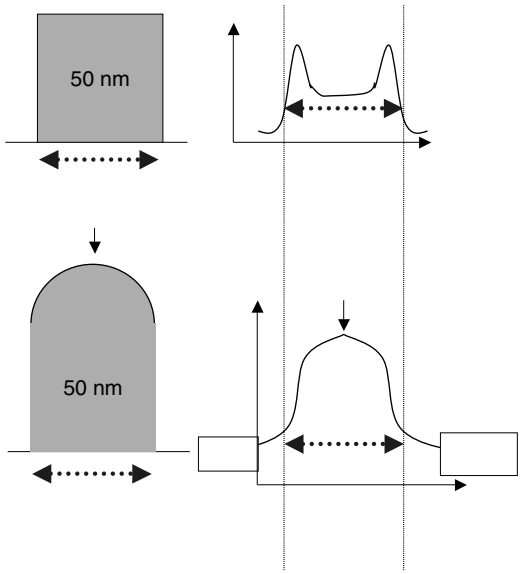


FIGURE 24.8 Impact of round topped resist lines on CD-SEM line scan.

Line shape information is critical to further processing and device performance. Formation of transistor source/drain and low dose drain features is done by implantation with the poly-silicon gate acting as a mask, and thus changes in shape can alter the implanted structure thus changing the transistor's electrical performance. Metal line resistivity can be effected by changes in line shape. Although modeling methods exist, the most used methods of controlling lineshape are the tilt beam CD-SEM and scatterometry. In Figure 24.9, we show the determination of line shape from tilt beam CD-SEM [17]. Other approaches include critical dimension atomic force microscope (CD-AFM) and destructive analysis using dual column FIB.

Marchman has shown the effect on CD value associated with the density of lines in the measurement area [12]. Additional charging is sometimes observed for dense lines. The effect varied with SEM supplier, and one SEM had a 20 nm change in measurement offset. The variation in offset with SEM model was attributed to different electron beam energies. The effect of line density can be minimized by selecting the appropriate electron beam voltage for each process step [12].

The effect of the electron beam on 193 nm KrF photoresist is well documented. Due to the large variety of photoresist and antireflective coating materials, the amount of shrinkage due to exposure to the electron beam is a function of materials as well as beam energy and dosage. Photoresist shrinkage can be minimized by using measurement recipes (procedures) that reduce exposure of the measured line or via to the electron beam and by optimizing the beam energy. Many different measurement schemes have been proposed for removing the effect of resist shrinkage from the data. These include measuring the feature from twice to several times before taking the final CD measurement. Recent reports indicate that photoresist shrinkage can be reduced or eliminated by using ultra-low ($\sim$100 eV) voltage CD-SEM [18].

Measurement and control of LER and line width roughness (LWR) are becoming more critical as the transistor gate length scales. Establishing a single definition of each of these quantities is difficult since standard definitions are under consideration. In that light, the definitions used in the 2003 ITRS were selected as a start. The LWR is 3σ of the CD over the spatial frequencies associated with the pitch and the drain extension [1]

$$\frac{1}{P} \leq \text{spatial frequences} \leq \frac{1}{0.5X_j},$$

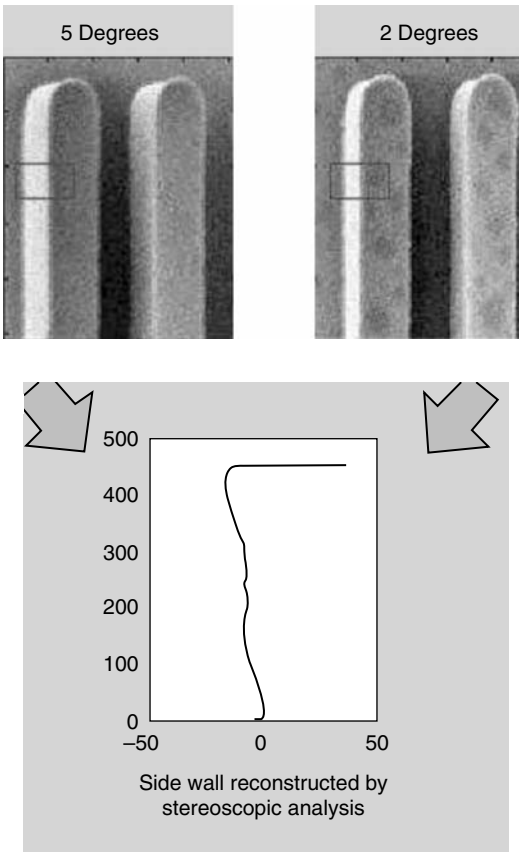


FIGURE 24.9 Tilt beam CD-SEM allows the reconstruction of lineshape figure courtesy Bob Burkhardt of applied materials.

The goal was to distinguish between line width and CD changes. Although this defines the spatial frequency range used to calculate LWR, it does not define the length of line that must be used to determine these frequencies. Given the highly statistical nature of roughness, sampling of line segments $\gg$ than the pitch will be necessary while sampling scan lines <5 nm apart [19].

24.2.1.2 CD-SEM Focus, Calibration, Tool Matching, and Manufacturability

A unified specification for CD-SEM technology has been developed for the sub 130 nm technology nodes [20]. This reference provides a broad overview of how one can judge the manufacturing compatibility of a CD-SEM. CD-SEM measurements are calibrated using reference materials, CD-AFM, comparison to cross-sectioned samples, and correlation to electrical linewidth measurements. We describe each method below.

SEM performance criteria include beam diameter, resolution, and sharpness [14]. The Fourier transform of a SEM image is a powerful means of monitoring sharpness [14]. The usefulness of this methodology has been demonstrated in a production environment [21].

CD-SEM magnification is calibrated using a reference material such as the NIST Standard Reference material 2090 [12]. Recently, NIST and Sandia National Laboratory have developed a CD reference material for physical and electrical CD measurement [22]. The test structure is both a microelectromechanical system, and an electrical test structure, and has specially etched, sharp sidewalls so that it can serve as a reference material for CD-SEM or CD-AFM. Since SEM measurements are strongly effected by the material (photoresist vs. poly silicon gate vs. TiN/Al/TiN vs. W/TiN), multiple reference samples

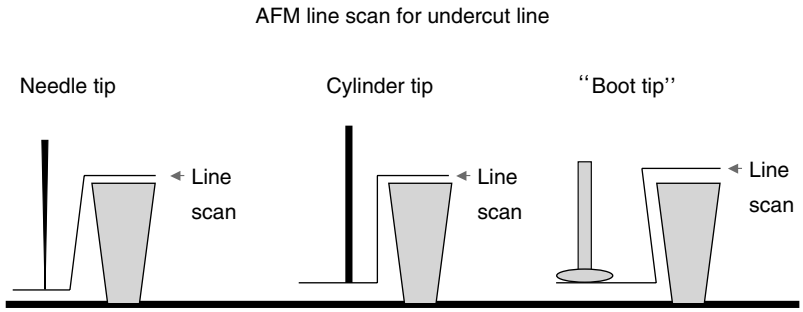


FIGURE 24.10 Probe tip shape greatly affects the observed atomic force microscope (AFM) image. There are samples where the cylindrical tips provide superior images. The cylindrical tip shape is expected to be the only mechanically stable tip shape for analysis of dense 100 nm features. Both etched fiber optic tips and carbon nanotubes are being investigated. (See from Dia, et al., *Nature*, 384, 147, 1996.)

provide optimum process control. A single reference material is considered appropriate for checking SEM drift and CD-AFM precision [22]. Commercialization of this reference material was not complete when this chapter was written.

Martin and Wickramasinghe pioneered the application of atomic force microscopy to lithographic metrology needs [23,24]. Griffith [25–27] and Marchman [28,29] have shown the utility of CD-AFM calibration of CD-SEM measurements. Ideally, CD-AFM measurements are independent of sample material. When the CD-AFM is itself well calibrated, it should provide the accuracy and precision required for CD-SEM calibration. An ASTM standard for characterizing probe tip shape is now available [30]. This is a critical part of maintaining measurement reproducibility. The CD-AFM measurement precision has been discussed in the literature [31]. Using "boot tips" (see Figure 24.10), the precision

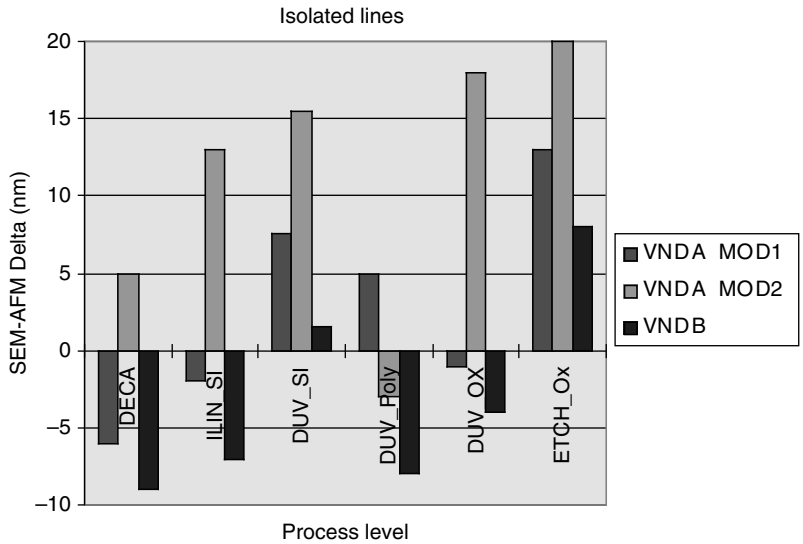


FIGURE 24.11 The effect of materials changes on CD-SEM offset was found to be a function of CD-SEM equipment model. Each SEM uses different primary beam voltages. Taken from reference 2.3 and used with the author's permission. Figure courtesy Hershel Marchman.

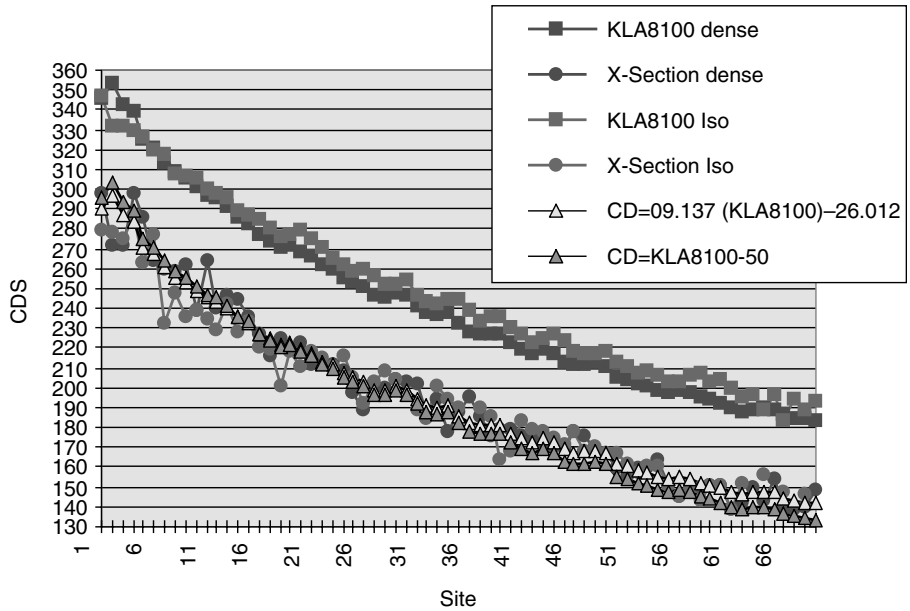


FIGURE 24.12 Calibration of CD-SEM using SEM cross-sectional data. Data for both isolated and dense lines show the same off-set over a broad range of linewidths. Figure courtesy Arnie Ford.

σ was found to be 2 nm for the top of the line and 1.5 nm in the middle of the line [31]. Height measurement precision was 0.5 nm and side wall angle was 0.21° for the left wall. In Figure 24.10, we show schematically how probe tip shape effects a line scan across an undercut line. This illustrates the relationship between the AFM image and the sample. The specific relationship will be different for each type of probe tip shape (e.g., sharp pointed tips vs. rounded boot tips). Calibration of CD-SEM requires that the CD-AFM have greater accuracy than the CD-SEM and appropriate precision. The different CD-SEM offsets associated with different materials found at subsequent lithographic process steps are shown in Figure 24.11.

Cross-sectional SEM is an effective method of calibrating CD-SEM. CD-SEM calibration is done across a range of linewidths as shown in Figure 24.12. Calibration data for both isolated and dense lines both show a constant offset from 350 to 130 nm linewidths.

In an effort to provide greater statistical significance, CD-SEM algorithms now allow for determination of linewidth for several lines in the image field. As the number of lines in this field increases, the average CD in the field provides identical information to scatterometry. Projecting into the future when it may be possible for both methods to sample the same number of lines in a measurement area, there will be one key difference between scatterometry and CD-SEM other than time per analysis area. The CD-SEM will be capable of providing the range and average of the local distribution of CD values. This development will be closely followed by the metrology community.

24.2.1.3 Scatterometry

Scatterometry refers to the use of scattered light to determine lineshape and CD. There are two main approaches to scatterometry, and Raymond's review is recommended [32]. In the single wavelength, multi-angle method, light is scattered from a grating test structure and collected at a series of angles [33]. The intensity maximums of diffracted light occur at angles that depend on the shape and width of the lines making up the grating structure. The polarization dependence of the data is often significant. This method requires specialized equipment for data collection. In the single-angle multi-wavelength method,

the reflected intensity versus wavelength also depends on linewidth and shape [32]. This is schematically shown in Figure 24.13. Commercial spectroscopic ellipsometers (SEs) found on in-line optical metrology systems can record the multi-wavelength single angle data. This method is also known as phase profilometry. This is schematically shown in Figure 24.14. In both methods, intensity of the reflected light is compared to a library of scattering patterns that simulate linewidths and feature shapes. Recently, computational improvements have allowed calculation of feature shape and CD without libraries. In Figure 24.13, we show both approaches to scatterometry.

In order to provide an overview of scatterometry, it is useful to start with the scattering or diffraction of light from a regularly spaced grating structure [32]. The intensity of the diffracted light occurs at specific angles given by the well known grating equation:

$$\sin\theta_{\mathrm{i}} + \sin\theta_n = n\frac{\lambda}{d}$$

where θ_{i} is the angle of incidence, θ_n is the angular location of the nth diffraction order, λ is the wavelength of incident light and d is the spatial period (pitch) of the structure [32]. Very careful analysis of the diffraction pattern allows one to directly determine lineshape and width and the thickness of patterned features and the some of the layers below [32]. In order to fully interpret the scattering patterns, a series of scattering patterns are simulated for a range of feature shapes and dimensions using the Rigorous Coupled Wave Approach [32]. The single wavelength, multi angle approach is known as "2-Θ" scatterometry [32].

The second method of scatterometry, single angle–multi wavelength, has two major variations. In one variation, an SE is used to measure the usual parameters Δ (called "Del") and Ψ using a diffraction grating as a target [33]. These parameters are defined below in the section on ellipsometry. The changes in Δ and Ψ from that expected for an unpatterned film stack allow determination of the average line shape and CD over the measured area. The second method is to use a perpendicular optical path for a reflectometer equipped with polarized light [34]. The second method requires that the parallel and perpendicularly polarized light be oriented relative to the lines in the diffraction grating.

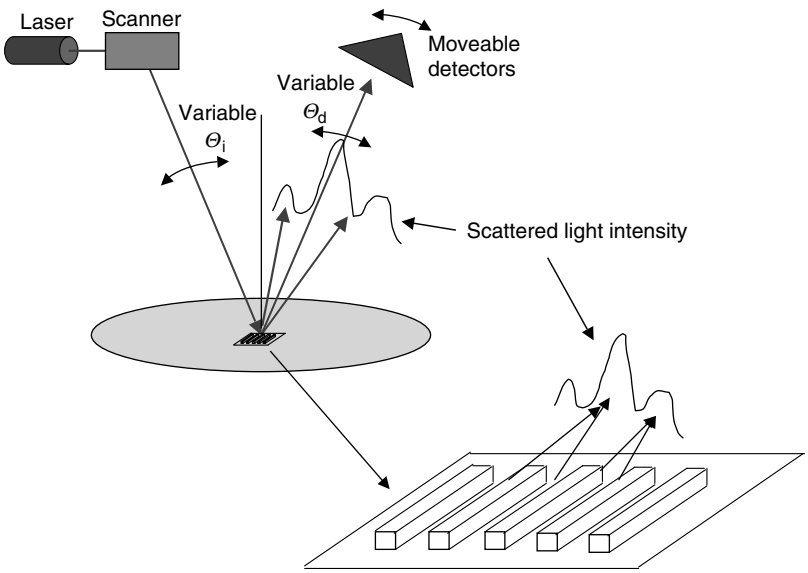


FIGURE 24.13 Diagram of single wavelength–multi angle scatterometry measurement of critical dimension (CD).

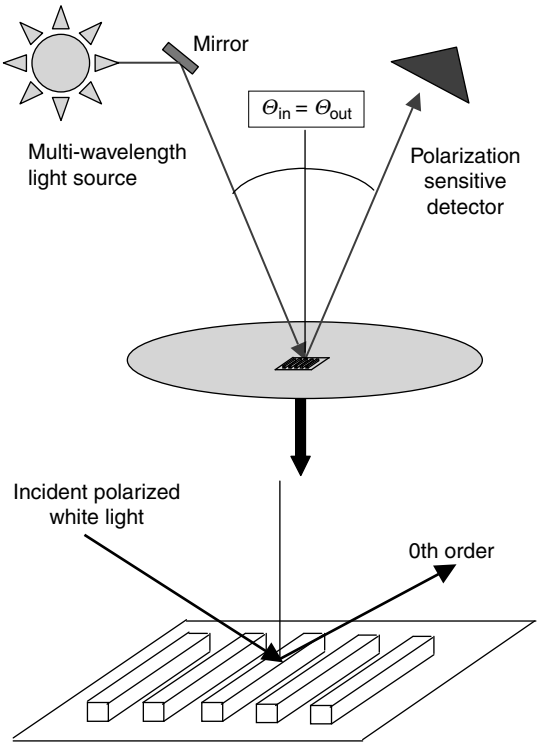


FIGURE 24.14 Diagram of multi wavelength–single angle scatterometry measurement of CD in-line spectroscopic ellipsometers can be used to measure CD.

As shown in Figure 24.14, the average line shape can measured with a great amount of detail with all forms of scatterometry. Since it is an optical method, all forms of scatterometry can be done quickly allowing more rapid throughput when compared to other methods. When comparing scatterometry to CD-SEM, it is important to note that scatterometry measures the average CD while CD-SEM measures one value from the average. Unpublished reports have proposed the need to average from 10 to >50 individual lines (measured by CD-SEM) to determine an average CD that tracks the average measured using scatterometry. There are many difficulties associated with direct comparison of CD-SEM and scatterometry including the need to determine exactly where along the line to compare (top vs. middle vs. bottom) [35].

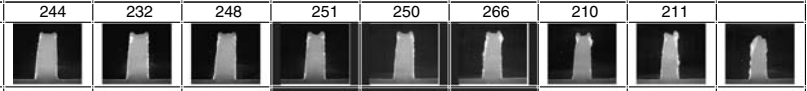
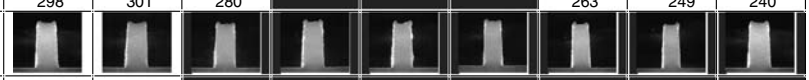
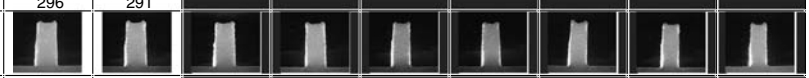
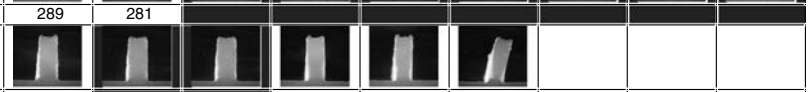
24.2.1.4 Application of CD Metrology

CD measurements can be applied at a number of steps during the fabrication of shallow trench isolation, the transistor gate electrode, and interconnect Damascene patterning. For example, CD can be measured in the patterned photoresist, after the resist is trimmed and/or after the poly-silicon is etched. The use of scatterometry combined with advanced process control methods for run to run process control at the etch steps are reported to provide much tighter control of the electrical properties of transistors such as the saturation drive current [36]. Since scatterometry determines film thickness too, the shape and depth of shallow trench isolation can be determined before oxide fill. At this time contact and via CD are done using SEM, and CD-SEM is also used during transistor fabrication. Thus, scatterometry is used to supplement CD-SEM by providing tighter process control at certain key steps.

Another function of CD metrology is to determine process tolerance ranges after a change in process recipe. The combination of CD-SEM with cross-sectional CD measurements enable control of sidewall

profile and linewidth. Scatterometry has been applied to control of lines during focus exposure, and dual column FIB has been applied to contact/via control. A recipe change is illustrated by an exposure vs. focus matrix for isolated and dense 250 nm lines respectively in Figure 24.15 [37]. The optimal process range is highlighted for isolated and dense lines. This data was taken after a double post-exposure bake process. In Figure 24.16 the effect of different post exposure bakes for this two bake step process are shown. CD-SEM data is correlated to cross-sectional data from the exposure matrix, and CD-SEM is then used for SPC.

250 nm isolated lines

Nominal feature: 225 nm Isolated	UVIIHS double PEB study 125°C/80s PEB#1+140°C/20s PEB#2								
Exp (mJ×cm^{-2})® focus (mm)$^{-}$	8.775	9.450	10.125	10.800	11.475	12.150	12.825	13.500	14.175
−0.1									
Feature size	244	232	248	251	250	266	210	211	
+0.1									
Feature size	298	301	280				263	249	240
+0.3									
Feature size	296	291							
+0.5									
Feature size	289	281							
+0.7									
Feature size	293			226	195	194			

(a)

250 nm nested lines

Nominal feature: 225 nm nested	UVIIHS double PEB study 125°C/80s PEB#1+140°C/20s PEB#2								
Exp (mJ×cm^{-2})® focus (mm)$^{-}$	8.775	9.450	10.125	10.800	11.475	12.150	12.825	13.500	14.175
−0.1									
Feature size		290	270	256	248	248	239	224	208
+0.1									
Feature size	305	280	272	263	254	251	240	226	220
+0.3									
Feature size	307	286	268	267	249	246	230	225	220
+0.5									
Feature size	325	303	277	279	260	248	242	234	224
+0.7									
Feature size		299	288	281	252	245	225	202	185

(b)

FIGURE 24.15 Cross-sectional CD analysis of a process recipe using (a) isolated and (b) dense lines figure courtesy John Petersen based on a figure from Ref. [37].

Summary table of double PEB test with UVIIHS

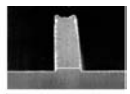
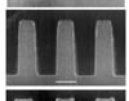
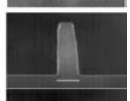
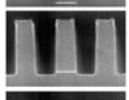
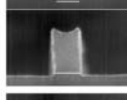
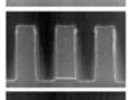
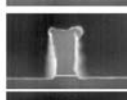
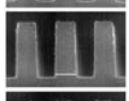
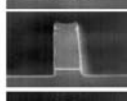
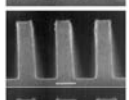
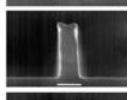
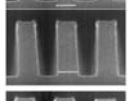
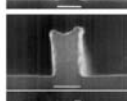
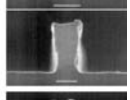
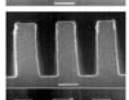
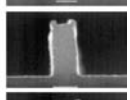
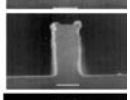
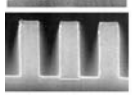
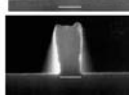
TEST number	PAB 90s °C	PEB#1 °C#1	PEB#1 s#1	PEB#2 °C#2	PEB#2 s#2	Nested lines	Isolated lines	Bias (nm)
1	130	125	80	140	20			44
2	130	140	10	125	80			0
3	130	125	80	150	20			64
4	130	150	10	125	80			32
5	130	125	40	140	20			39
6	130	140	10	125	40			15
7	130	125	40	150	20			47
8	130	150	10	125	40			49
9	130	125	60	145	15			38
10	130	145	15	125	60			45
Reference	130	140	90	None	None			83

FIGURE 24.16 Example of process set-up illustrated using a double bake resist process. Figure courtesy John Petersen.

24.2.1.5 Electrical CD Measurement

There are two approaches to electrical measurement of line width. The effective transistor gate length can be determined from the transconductance of the transistor. Another approach is to measure the electrical properties of a test structure such as the cross-bridge resistor test structure, which is shown in Figure 24.17. The electrical line width is determined by first measuring the sheet resistance of the van der Pauw resistor and then measuring the resistance of the bridge resistor. To be more specific, we refer to Figure 24.17 and the description of Ref. [38]. A current I is placed between pads 4 and 3 and the voltage V_1^+ is measured between pads 5 and 2. The polarity of the current is then reversed and voltage V_1^- is measured. Then the current I is placed between pads 3 and 2 and the voltage V_2^+ is measured between pads 4 and 5. The polarity is again reversed and voltage V_2^- is measured. The sheet resistance is calculated using [38]:

$$R_s = \frac{\pi(|V_1^+| + |V_1^-| + |V_2^+| + |V_2^-|)}{4I \ln 2}$$

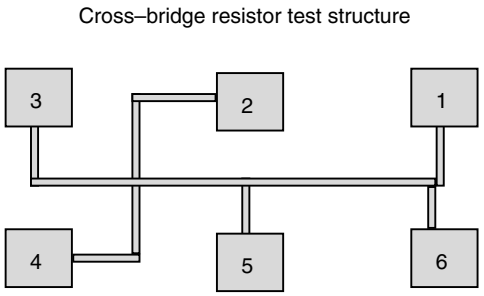


FIGURE 24.17 Cross-bridge resistor electrical test structure for CD measurement.

The bridge resistance is calculated from voltage measurements between pads 5 and 6 while a current is placed between pads 1 and 3. V_b^+ is the voltage for current placed between pads 3 and 1, etc., From this the electrical linewidth can be calculated using:

$$R_b = \frac{(|V_b^+| + |V_b^-|)}{2I_b} \text{line width } W = \frac{R_b L}{R_s}$$

L is the length between pads 5 and 6.

In-line electrical CD measurements of test structures on product wafers have been used to control CD during volume manufacture. Automated, in-line electrical probers equipped with camera and pattern recognition are combined with automated electrical measurement systems. A high correlation of CD-SEM and electrical CD measurements has been observed. In one study, correlation coefficients of at least 0.99 were found for both poly-silicon lines and aluminum lines coated with anti-reflective coating (ARC), the slope of both lines indicated that the CD values from the CD-SEM were wider than the electrical measurements [39].

The transconductance, g_m, of a transistor provides another measure of electrical gate length, L_{eff}. Transconductance is defined as the derivative of the drain current with respect to the potential difference between the gate and source, V_{GS}, at a constant potential difference between the drain and source [40].

$$g_m = \frac{dI_d}{dV_{GS}} | V_{DS}$$

Plots of the drain current vs. V_{GS} show two different slopes, one at low V_{GS}, and one after the transistor reaches saturation voltage, V_{SAT}. Therefore, an ideal transistor will show two relationships between g_m and L_{eff} as follows:

$$g_m = \{\beta V_{DS} \text{ for } V_{DS} \leq V_{SAT} \text{ or } \beta V_{GT} \text{ for } V_{DS} > V_{SAT}\}$$

Here, $\beta \propto 1/L_{eff}$ [40]. This type of CD measurement can be used to separate chips according to expected circuit speed due to line-width-induced gate delay.

24.2.2 Overlay Process Control

Overlay is the term used to describe the registration of the patterned structure in a layer with the patterned structure in the subsequent layer [41,42]. Most overlay metrology is done using an optical system that automatically evaluates how far from center the target pattern in the top layer is from the center of the target pattern in the layer below. Every lithographically patterned layer requires overlay control. Tight control of overlay is a critical part of IC manufacture of the three dimensional structures

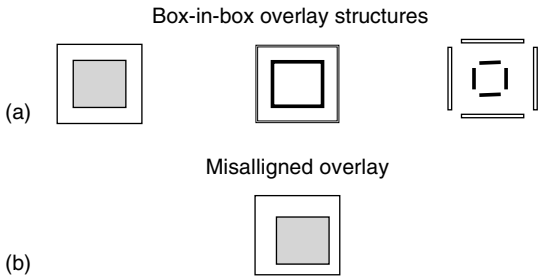


FIGURE 24.18 Typical overlay patterns. Figure suggested by Chris Nelson.

that form the integrated circuit. Process levels with the smallest IC structures such as the transistor gate drive the metrology tool requirements for overlay.

Optical overlay metrology requires test patterns that provide high contrast through the dielectric layer that separates the gate level from the first metal level and subsequently the metal levels from the metal level below. Overlay is used after each lithographic mask step except the first. Typical overlay patterns are shown in Figure 24.18a, and a misaligned box target is shown in Figure 24.18b. A cross-sectional view of an overlay box in box structure and the line structures is shown in Figure 24.19. In Figure 24.20, we show a block diagram of a lithographic stepper and misalignment errors associated with distance between reticle and lens, reticle flatness, and wafer rotation. Overlay must be checked inside each die (intrafield) and from die to die (interfield). Interfield comparisons are done using overlay patterns from the center of the die. In Figure 24.21, we show a wafer with a four point check of interfield errors. The lithographic stepper or step-and-scan tool must correct for translation errors, scaling (different size translation errors across the wafer), orthoginality (centering of overlay pattern), and wafer rotation (see Refs. [35,36]). In practice, a monitor wafer with overlay artifacts is run each day to check inter and intra-field errors.

Advanced interconnect processes will challenge overlay process control. Overlay of chemical mechanical polishing (CMP) process steps is made difficult by the fuzzy image observed for the buried box structure. Separating the overlay error due to the stepper from the error in the overlay measurement is often difficult. The origin of the fuzzy image is shown in Figure 24.22. Overlay of Damascene processes is also difficult.

The overlay metrology tool consists of a specially designed optical microscope, a camera with a digital detector, and the illumination source [41,42]

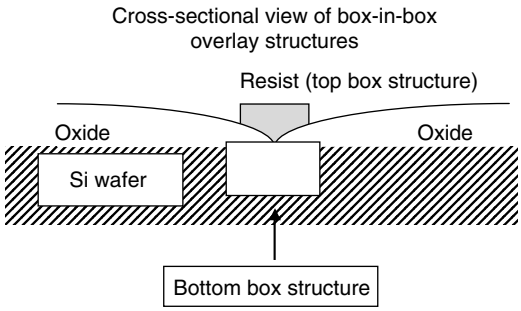


FIGURE 24.19 Cross-sectional view of overlay box and line structures. Figure suggested by Chris Nelson.

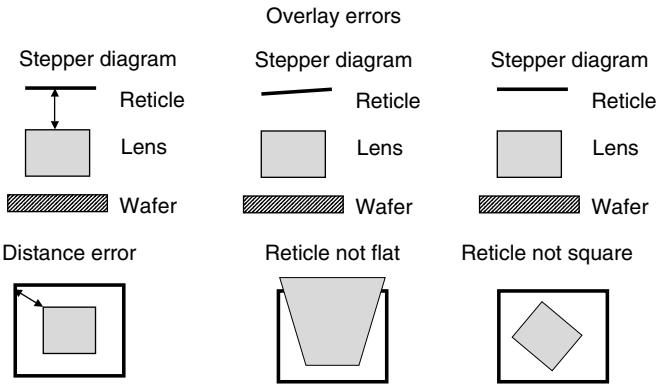


FIGURE 24.20 Relationship between lithographic stepper and misalignment errors. Figure suggested by Chris Nelson.

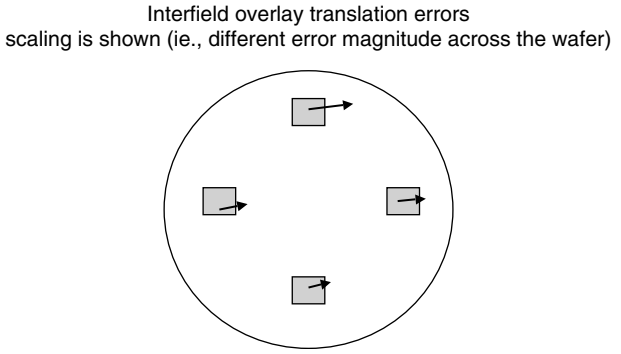


FIGURE 24.21 Four point check of interfield errors. Figure suggested by Chris Nelson.

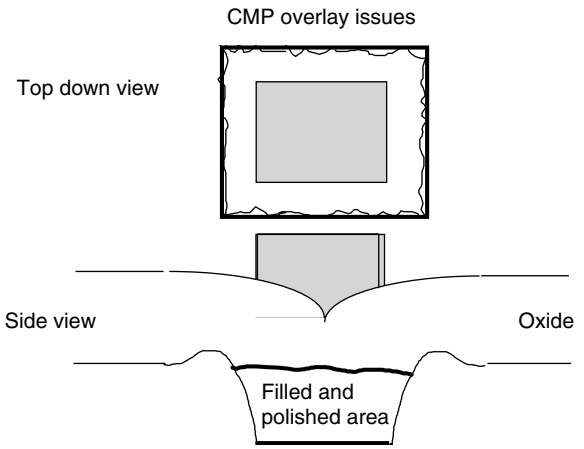


FIGURE 24.22 Chemical mechanical polishing (CMP) issues for overlay measurement. Separation of overlay error from measurement error is difficult due to CMP induced edge fuzziness. Figure suggested by Chris Nelson.

24.3 Metrology for Front End Processes

Transistors continue to evolve rapidly. New materials will enable current CMOS, and new transistor designs such as the FINFET are expected to extend CMOS beyond the 22 nm node. One critical challenge facing the industry is scaling the properties of transistors such as switching speed. The properties of a CMOS inverter are often used to illustrate the relationship between the transistor's saturation drive current, I_{dsat}, and the gate delay, $\tau = C_{load}V_{DD}/I_{dsat}$. The relationship between transistor characteristics and saturation drive current is a function of gate length. In long channel devices (CD > 100 nm), saturation drive current was directly proportional to carrier mobility and inversely proportional to gate length and dielectric thickness. For ultra-short channel devices, the saturation drive current is proportional to the saturation velocity of the carriers, the transistor width, the capacitance, and on voltage. The electrical properties will transition between long and short channel behavior. Thus, 65–32 nm node transistors will continue to use stress to improve carrier mobility, but the impact on saturation drive current will decrease.

Fabrication of transistors requires a number of different processes including formation of isolation structures such as shallow trench isolation, growth of the gate dielectric, doping, and lithographic patterning of gate electrodes. A summary of the metrology steps found in typical from end processing are illustrated in Figure 24.23. In this section, the topics of gate stack film thickness and electrical measurements and dopant dose and junction measurements are discussed. A great variety of process methods are used to induce stress in the transistor channel, and each one requires a different approach to process control. Stress measurement is also reviewed. Gate dielectric thickness is routinely measured on patterned wafers using either multiple angle or SE. A recently introduced in-line x-ray photoelectron spectroscopy (XPS) based method is briefly described along with electron beam induced x-ray fluorescence (XRF). These methods have the advantage of determining nitrogen content. Electrical characterization of gate dielectrics includes non-contact corona based methods and traditional capacitance–voltage (C–V) and current–voltage (I–V) measurements. The ability of the non-contact

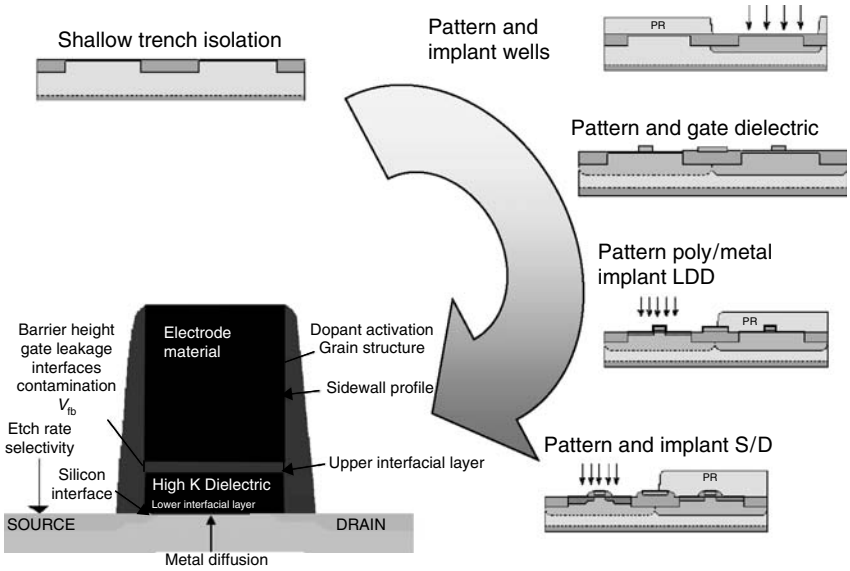


FIGURE 24.23 Typical metrology steps used during transistor fabrication. Control of shallow trench isolation, implants for well formation, gate dielectric formation, low dose drain implants, and source/drain implants all require statistically significant measurements for process control.

methods to measure on patterned wafers is continuously evolving. Dopant dose measurements often require special test wafers, while junction measurements can be done on patterned wafers in special test areas. The use of metal gate electrodes instead of poly-silicon is predicted to occur below the 65 nm technology node, and it is briefly discussed here.

The dielectric thickness is determined using a model of the optical properties of the film structure and the observed thickness value can depend on the size of the analysis area and the specifics of interface properties used in the model. Silicon oxynitride thermal films less than 2 nm in thickness and the contribution of the interface to optical and electrical properties is significant. The impact of adsorption of ambient contamination on the dielectric surface has required desorption of these films before optical thickness measurements. Electrical capacitance based measurement of effective dielectric thickness can be done using a fabricated capacitor structure or with equipment that makes the top of the capacitor through corona discharge or with the elastic metal tip approach.

24.3.1 Ellipsometric Measurement of Gate Dielectric Film Thickness

Ellipsometric measurement of film thickness is based on the change in polarization of light after reflection from the dielectric on the wafer surface. If the optical constants of the dielectric are known, the thickness can be determined using a model of the optical structure of the dielectric film on a silicon wafer. The theory of ellipsometric measurement is left to references such as Tompkins [43] and Jellison [44]. In order to better describe ellipsometric data, a brief summary is presented below.

24.3.1.1 The Theory of Ellipsometry

A brief description of elliptically polarized light and the phenomena of reflection and refraction facilitates the discussion of ellipsometry given below [43,44]. In Figure 24.24, a wave of light that is linearly polarized parallel to the plane of incidence to the sample surface is shown reflecting at an angle ϕ. This wave is designated a "p" wave. Light that is polarized perpendicular to the plane of incidence is designated as "s" waves. When "p" and "s" waves are combined slightly out of phase, the light beam is elliptically polarized. If the phase difference is 90°, the light is circularly polarized. In Figure 24.25, light is shown reflecting from an infinitely thick sample with complex index of refraction $N_2 = n - ik$ where k is the extinction coefficient. The extinction coefficient is related to the adsorption coefficient of light, α, through the following relationship:

$$k = \frac{\lambda}{4\pi}\alpha \text{ and } I(z) = I_0 e^{-\alpha z}$$

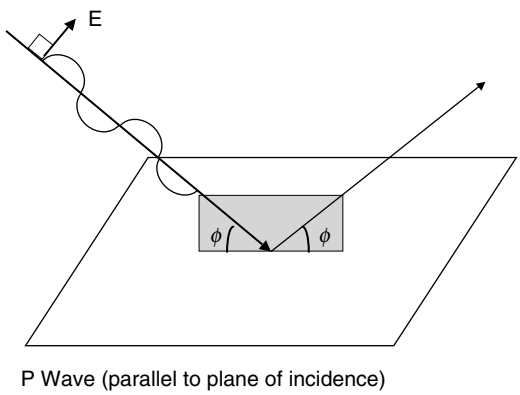


FIGURE 24.24 Reflection of "p" polarized light from a sample surface.

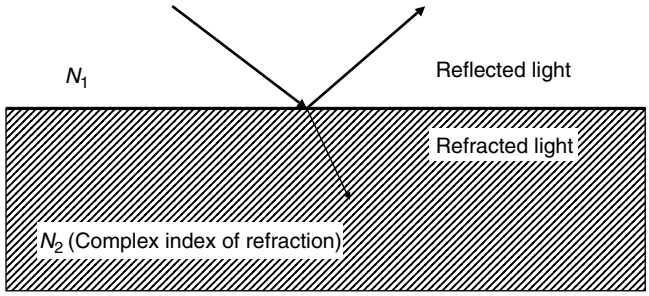


FIGURE 24.25 Refracted vs. reflected light.

$I(z)$ is the intensity of light at a depth z below the sample surface, and I_0 is the initial intensity of light that enters the sample.

Since ellipsometric measurement is based on the change in the elliptical polarization after reflection, we must relate the reflection coefficient to the film thickness. The Fresnel reflection coefficients for reflection from an infinitely thick sample for polarizations both parallel and perpendicular to the incident plane are:

$$r_{12}^{\mathrm{P}} = \frac{N_2 \cos\phi_1 - N_1 \cos\phi_2}{N_2 \cos\phi_1 + N_1 \cos\phi_2} \qquad r_{12}^{\mathrm{s}} = \frac{N_1 \cos\phi_1 - N_2 \cos\phi_2}{N_1 \cos\phi_1 + N_2 \cos\phi_2}$$

This leads to the complex, total reflection coefficients for a multi-layer stack:

$$R^{\mathrm{P}} = \frac{r_{12}^{\mathrm{P}} + r_{23}^{\mathrm{P}} \mathrm{e}^{-i2\beta}}{1 + r_{12}^{\mathrm{P}} r_{23}^{\mathrm{P}} \mathrm{e}^{-i2\beta}} \qquad R^{\mathrm{S}} = \frac{r_{12}^{\mathrm{s}} + r_{23}^{\mathrm{s}} \mathrm{e}^{-i2\beta}}{1 + r_{12}^{\mathrm{s}} r_{23}^{\mathrm{s}} \mathrm{e}^{-i2\beta}}$$

In Figure 24.26, we show the multiple reflections described by the Fresnel coefficients used above. The phase thickness β is related to the physical thickness by

$$\beta = 2\pi \frac{d}{\lambda} N_2 \cos\phi_2$$

The approximate nature of this model is illustrated by the transmission electron micrograph of a 3 nm thermal oxide film shown in Figure 24.27. The interface between the oxide and silicon substrate is not

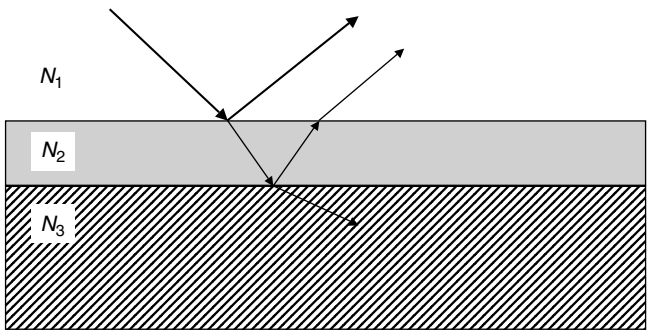


FIGURE 24.26 Reflection of light from a layered sample.

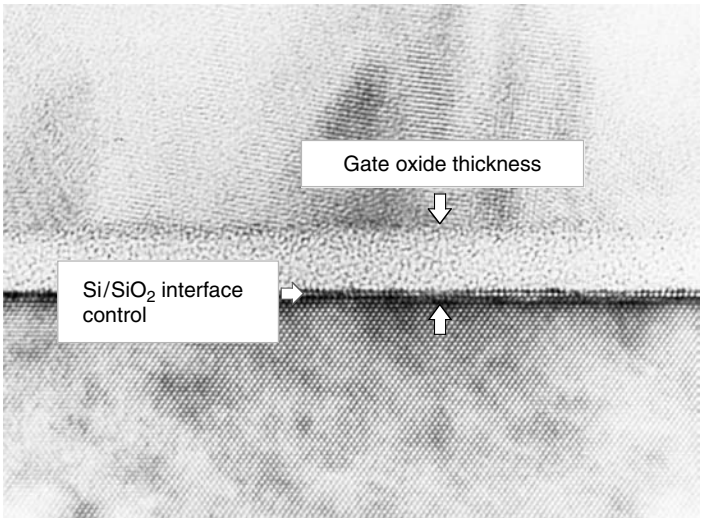


FIGURE 24.27 Transmission electron micrograph of a ~3 nm silicon dioxide gate dielectric. This image illustrates the lack of a sharp interface between the oxide and substrate. Figure courtesy Bob McDonald.

sharp. A complete model for measurement of gate dielectric thickness includes the optical effect of the interfacial region. In addition, local thickness variations across the area and the size of a sub micron gate will be averaged in typical commercial measurement systems. Averaging over a large area results in sub 0.1 nm precision. Local thickness variations can result in threshold voltage variations. Total reflection coefficients for any model and optical constants of the oxide layer can be programmed into commercial, in-line systems.

The ellipsometer actually measures the parameters Δ (called Del) and Ψ. Del is the phase difference after reflection, $\Delta_1 - \Delta_2$. $\Psi$is defined by the following equation:

$$\tan \Psi = \frac{|R^{\mathrm{p}}|}{|R^{\mathrm{s}}|}$$

The fundamental equation for ellipsometry is:

$$\rho = \frac{R^{\mathrm{p}}}{R^{\mathrm{s}}} = \tan \Psi e^{i\Delta}$$

In Figure 24.28, we show how Δ and Ψ change with thickness for a SiO_2 film. After a thickness of 283.2 nm, the value of Δ and Ψ will repeat.

24.3.1.2 Multi-Wavelength and Spectroscopic Ellipsometry

Gate dielectric thickness measurement requires exceedingly good precision. This has motivated the use of optical metrology systems that have multiple optical paths. These systems include single wavelength ellipsometers (SWE) for optimum precision for thin films, spectroscopic reflectometers for thicker films, and SEs for complicated films and film stacks. Some commercial ellipsometers use one of 4 single wavelengths for gate dielectric control. The commercial 4-wavelength systems use optics that simultaneously measures over a range of angles. Film thickness is determined in SEs by fitting Δ and Ψ data taken at many wavelengths to a values calculated from an optical model of the sample.

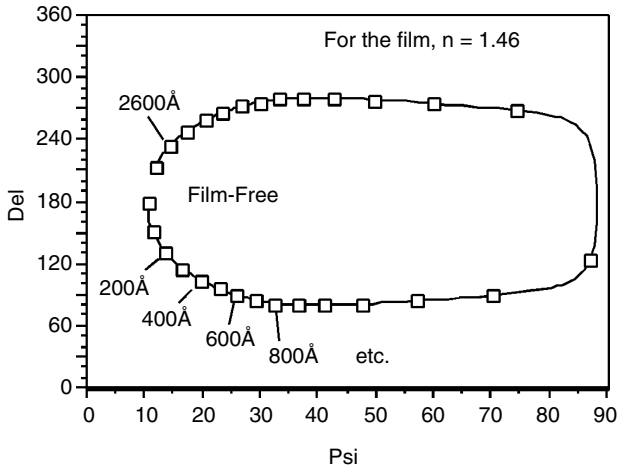


FIGURE 24.28 Trajectory of Del and psi for silicon dioxide. This figure shows that psi and Del trajectory will repeat after 283.2 nm thus making film thickness determination ambiguous for thicker films. This is one of the reasons that multiple wavelengths are used in many film thickness tools.

24.3.1.3 Ellipsometric Systems

In Figure 24.29a–c, we show a block diagrams of multiwavelength and SEs. A rotating polarizer type ellipsometer is shown in Figure 24.29a since it is typical of the design used in many commercial systems. There are a variety of system designs, some of which incorporate a focus system that averages over a spread of angles as shown in Figure 24.29b and c. Precision requirements for ultra-thin gate dielectric measurement may drive the use of non-focused systems. Detailed technical descriptions of commercial ellipsometers are available from the technical literature [45–47].

24.3.1.4 Optical Models for Thin Gate Films

The traditional model of an oxide layer on a silicon substrate is that of a single, flat layer that has a sharp interface with the silicon below. This layer has also been referred to as a slab. Although an interfacial layer with different physical and chemical properties is present after SiO_2 or SiO_xN_y is grown, a comparison of the precision of optical models for SE systems has shown that including the interfacial layer usually results in a larger (worst) precision than modeling the entire film as a single layer (or slab) [48]. Attempts at modeling using an interfacial layer composed of a Bruggerman Effective Medium Approximation (BEMA) model, provides a better Goodness-Of-Fit to the experimental data but a larger precision [48]. These effects are due to too much of correlation between fit variables and usually result in unrealistic values. Alternative single layer models (without an interfacial layer) are used that provide better precision values without sacrificing the Goodness-Of-Fit. Since SWE provides improved (i.e., smaller values) precision for these thin films, the optical constant for a single SiO_2 or SiO_xN_y layer is used. This approach assumes that the nitrogen concentration and depth profile remain uniform across the wafer and from wafer to wafer. The accuracy of oxynitride film thickness depends on the accuracy of the optical constants for that film. According to the ITRS, if the gate dielectric EOT is 1.0 nm thick and the process tolerance is 5% for 3σ (process variation), then $P/T = 10\% = 6\sigma/(0.1\text{ nm})$ which gives a measurement variation $3\sigma = 0.004$ nm. This precision can be achieved by desorbing the surface contamination layer that adsorbs on the wafer surface prior to measurement. A number of approaches have been made commercially available including use of infra-red laser irradiation and thermal desorption [49].

High-κ dielectric materials are being investigated as potential replacements for silicon oxynitride. The optical properties of these films are highly variable according to the method of deposition and the process

conditions. Further, there are no well known tabulated form of optical constants available for films such as hafnium oxide or silicate. Therefore, to accurately model these types of films, a parameterized model is needed that gives the optical properties of the corresponding film for the necessary wavelengths. To date, the Tauc–Lorentz model has proven a suitable choice. This model combines the classical Lorentz oscillator and the Tauc expression for the band gap of amorphous materials to give a parameterized function that models the imaginary part of the dielectric function [49]. However, two problems appear to still exist with this model: (1) no current manufacturing capable SE is capable of implementing this type of optical model. (2) There is too little optical contrast between the hik layer and the interfacial layer, the Tauc–Lorentz model is unable to distinguish the two layers and usually combines the two thicknesses as one. Better models are being developed.

24.3.1.5 Resolution for Ultra Thin SiO_2 and Alternate Dielectrics

For an SWE using a laser operating at 632.8 nm, a change in thermal SiO_2 thickness of 0.1 nm results in a change in $\Delta \sim 0.25^\circ$ for films that are between 0 and 10 nm in thickness. The change in Ψ at a single wavelength is not a straight line function between 0 and 4 nm thickness (3.1). A commercial ellipsometer

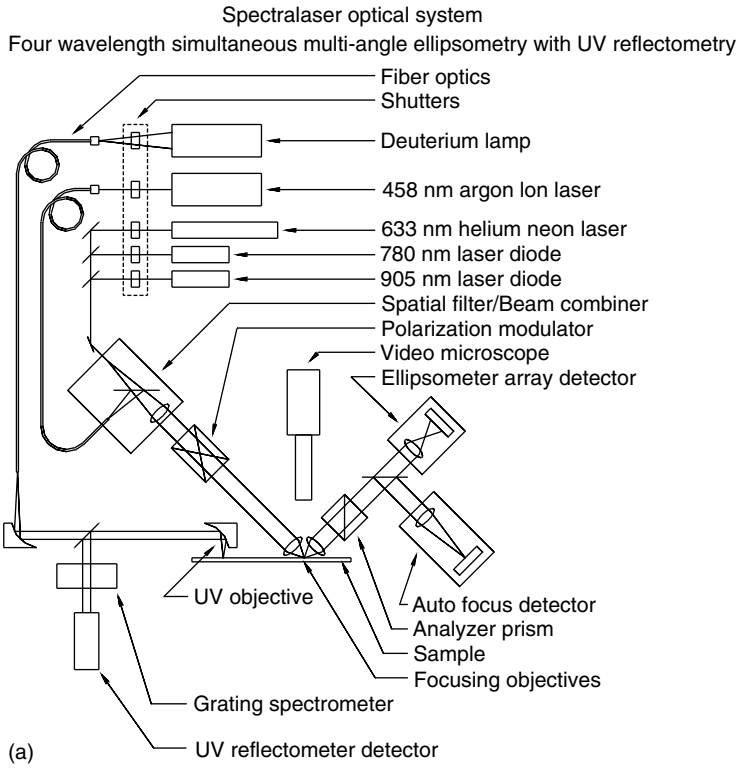
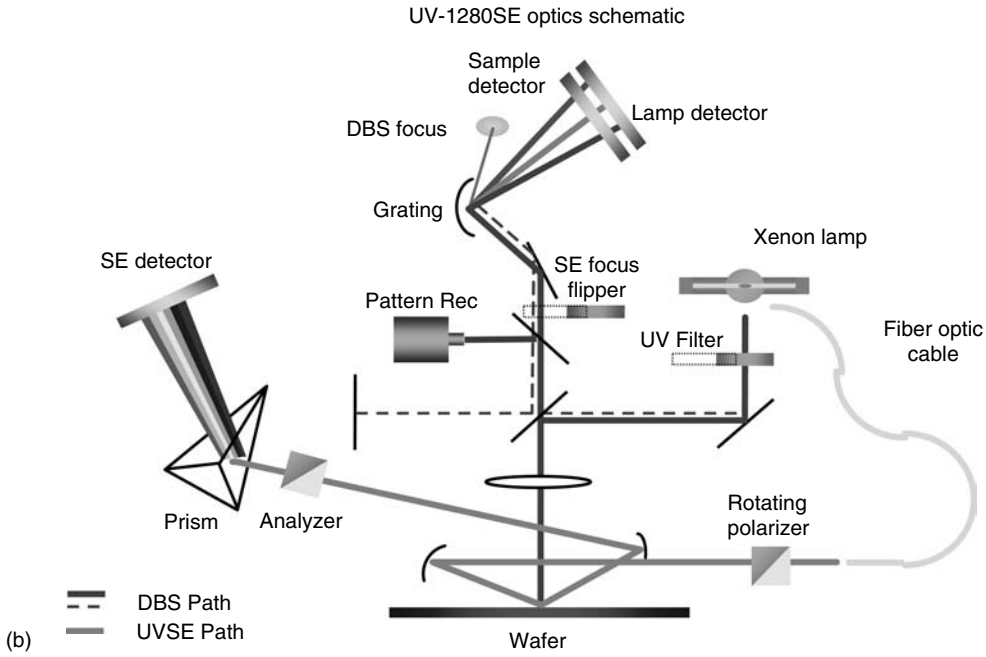
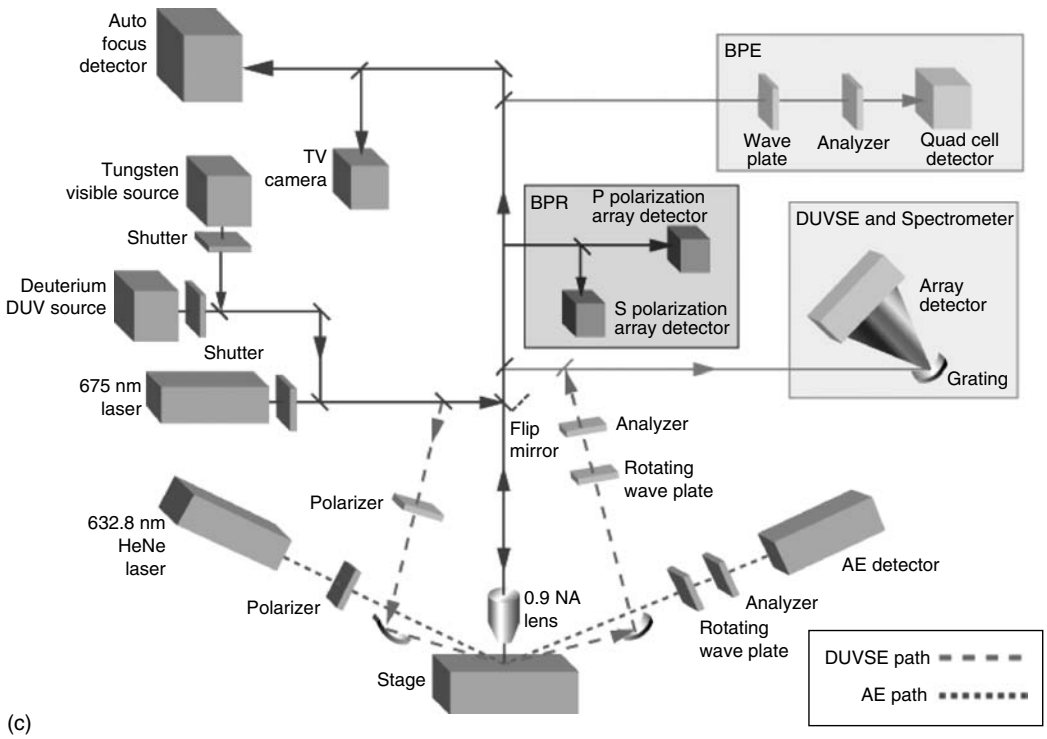


FIGURE 24.29 Block diagrams of several commercial ellipsometer designs. (a) Four wavelength, focusing, multiangle ellipsometer system made by Rudolph Technologies, Inc. Gate oxide thickness measurements can be made using only the 632.8 nm wavelength from the HENe laser. Figure courtesy J. Sullivan, Rudolph Technologies. (b) UV-1280SE spectroscopic ellipsometer system made by KLA-Tencor. The light source is a broad band Xenon lamp, and the ellipsometer is a rotating polarizer type. Figure provided by J.J. Estabil, KLA-Tencor. (c) Thermawave Optiprobe 5240 combined absolute ellipsometer (632.8 nm HeNe laser), spectroscopic ellipsometer (450–840 nm [W halogen lamp], and 210–450 nm [deuterium deep UV source]), beam profile reflectometer, beam profile ellipsometer, deep UV spectrometer (190–840 nm). Figure courtesy W. Lee Smith, Thermawave.



(b)



(c)

FIGURE 24.29 (*continued*)

is capable of measuring changes in Δ and Ψ of $<0.01°$. Several groups have discussed the sources of measurement error which effect accuracy and precision [49–53]. An error of 0.05° in incident angle near 70° results in a small error in Δ of $<0.02°$ and in $\Psi$of $\sim 0.1°$ when $\lambda = 632.8$ nm [50]. Multiple wavelengths provide a method of averaging data that helps overcome errors.

The Drude approximation predicts a linear change in Δ vs. thickness for thin (<10 nm), non-adsorbing (imaginary part of refractive index is small or zero) films of different refractive indices. Data for several films such as silicon dioxide and silicon nitride show that the parameter Δ (at $\lambda = 632.8$ and 70° incident angle) increases as refractive index increases [43]. The real part of the refractive index of silicon nitride and titanium dioxide is larger than that of SiO_2. Therefore, the resolution of ellipsometry to changes in a single layer film of these materials on a silicon substrate would be better than for SiO_2. The refractive index of non-isotropic materials such as single crystal TiO_2 must be accounted for, and an averaging procedure would provide the needed information for a film having randomly oriented grains. Since grain texture is process dependent, the optical constants of films made using new processes must be checked before using ellipsometry for process control. Thin dielectrics below 100A, Ψhas been shown to be almost independent of thickness. Alternatively, Δ changes almost linearly with thickness. Furthermore, $\Delta \sim 0$, for very thin dielectrics. Therefore, the sensitivity for thin dielectrics is with Δ. The SWE that use a Rotating Compensator has an intensity output proportional to Sin Δ. Considering the small angle approximation (Sin $\Delta \sim \Delta$), any resolution of thickness changes for thin dielectrics, will be detected using SWE with a Rotating Compensator.

24.3.1.6 Poly Si Thickness

The issues associated with single wavelength measurement of poly-crystalline silicon thickness have been discussed in the literature [43]. Optical wavelength ellipsometry is hampered by the fact that silicon is adsorbant in this range, while silicon is transparent in the infra-red region. Recently, in-line ellipsometer systems have been equipped with infra-red wavelength capability. The biggest issue is that the refractive index changes with micro-crystallinity [43]. It is also known that the micro-crystallinity of poly-silicon or a-Si can change across a wafer. Poly-silicon and a-Si thickness measurement is routinely done with commercial systems after careful consideration of the above issues.

24.3.2 Electrical Measurement of Gate Oxide Thickness

This section briefly discusses the measurement of the effective dielectric thickness using capacitors or transistors. Several references which provide details of capacitance measurements are recommended as supplementary reading [54–60]. In addition to these reviews, theoretical corrections have been used to extend capacitance measurements to 2 nm SiO_2 [60].

It is useful to discuss capacitance–voltage measurement in order to better compare optical and electrical measurement methods. The effective dielectric thickness refers to the thickness of the region that acts as a dielectric in the capacitor or transistor. The electrically measured thickness can be different from that measured optically when there is depletion of carriers in the poly-silicon above and/or in the silicon below the gate dielectric. In Figure 24.30a, a plot of capacitance vs. voltage (called a *C–V* curve or plot) for an ideal capacitor illustrates the response of a SiO_2 film that is greater than 4 nm on a uniformly doped *p*-type substrate [54–59]. The assumption is that the poly-silicon gate and the uniformly doped *p*-type silicon both act as metal plates in a perfect capacitor when the applied voltage on the gate is negative. Band bending causes positive charge accumulation at the surface of the *p*-type silicon just as one expects positive charge buildup on a metal plate capacitor as illustrated in Figure 24.30b. The capacitance of P-channel metal-oxide semiconductor (PMOS) structure drops as the voltage moves from negative toward zero. In a defect free structure the valence and conduction bands in the substrate are flat at an applied voltage equal to the flat band voltage. The flat band voltage for two "metal" plates having the same work function is zero as discussed in the example shown in Ref. [55]. The equilibration of the Fermi levels (between the doped poly silicon gate and uniformly doped substrate) results in band bending at zero gate voltage and ideal flat band voltage $V_{fb} = (\phi_m - X_s - E_c + E_F)/q$. ϕ_m is the work function of the

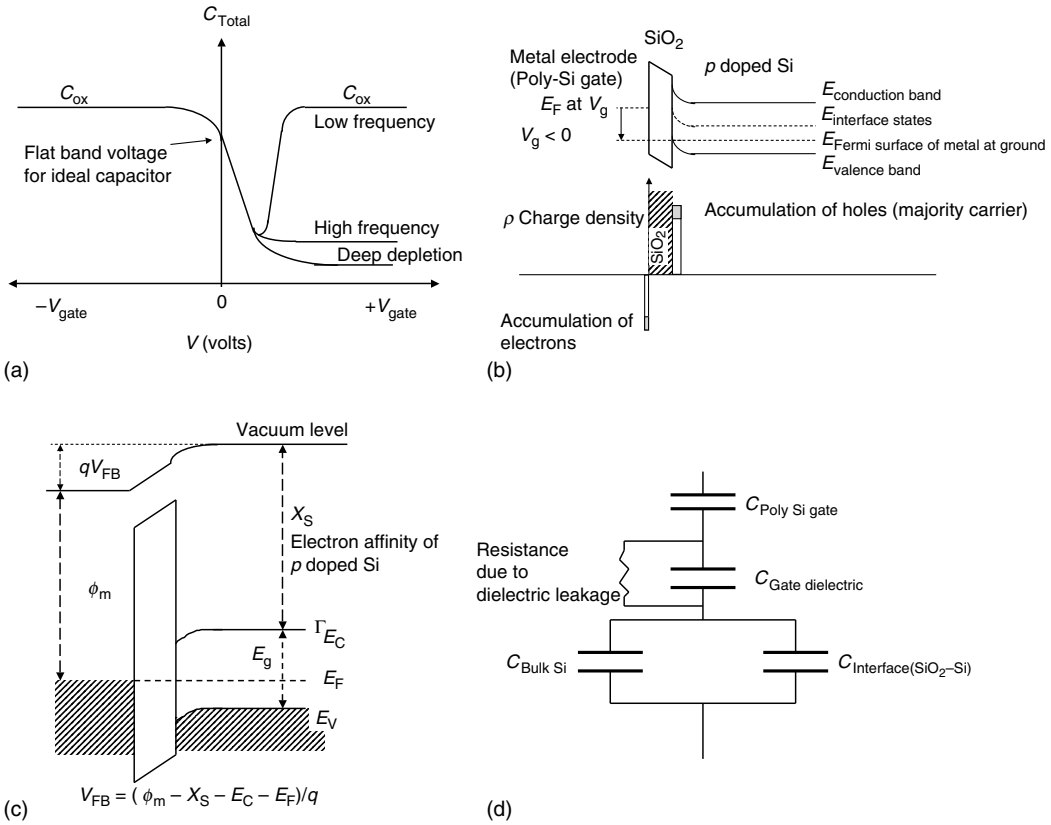


FIGURE 24.30 Capacitance–voltage measurement of oxide thickness. (a) Typical *C–V* data is shown. The oxide thickness is determined from the constant capacitance at negative voltage for *p*-type silicon substrates. (b) The band bending diagram and charge density plot for gate electrode and p-doped substrate show the charge accumulation at the p-substrate–gate oxide interface. (c) The origin of the flat band voltage. The energy band diagram for the poly-silicon/gate oxide/p-doped substrate illustrates band bending at zero applied voltage. (d) The equivalent circuit diagram shows the various contributions to the capacitance of the transistor gate.

metal, X_s is the electron affinity of the doped silicon substrate, E_c is the conduction band edge of the doped silicon substrate, E_F is the Fermi level, and q is the electric charge. A band diagram of the poly-silicon/SiO_2/"p" type Si system based on Ref. [60] is shown in Figure 24.30c [60]. The *C–V* data is obtained by sweeping the voltage at either low or high frequency. This frequency has a large effect on the capacitance observed at positive voltage as shown in Figure 24.30a. Series and parallel capacitors have been used to describe the capacitance of the entire system as shown in Figure 24.30d [59]. The capacitance of the oxide is in series with charge in the semiconductor and in the poly-silicon gate. The charge in the semiconductor is in parallel with the interface charge, and Schroder divides the charge in the semiconductor into the hole accumulation charge "C_p", the space-charge region bulk charge, and the electron inversion charge which is not shown in Figure 24.30d. When the accumulated charge in the *p*-type silicon is very large, and the C_p is considered to be shorted and thus it acts as a perfect metal plate in an ideal metal-oxide semiconductor (MOS) capacitor. For thicker oxides >4 nm, the gate electrode is considered to be a perfect metal capacitor plate. Using these assumptions, the effective thickness can be calculated using the following relationship, $C_{ox} = \varepsilon_{ox} A/d_{ox}$ [55–60]. C_{ox} is the maximum capacitance at negative applied gate voltage for PMOS, d is the effective dielectric thickness, A is the area of the capacitor, and ε_{ox} is dielectric constant (real part of dielectric constant). Defects in the oxide layer result

in the trapping of charge in the oxide layer [54–60]. This shifts the flat band voltage from its ideal value, and allows the quality of the gate dielectric to be monitored by the value of the flat band voltage.

Several studies have shown an excellent correlation between C_{ox} and ellipsometric oxide thickness. By this we mean that a plot of the uncorrected electrical thickness vs. physical thickness is linear. The electrical and physical measurements will both have contributions from the interface between the gate oxide and silicon substrate. The dielectric constant of the interfacial region is different from bulk SiO_2. This is not corrected for in the electrical determination of thickness. For the very thin oxides <2 nm, *C–V* behavior is not ideal, i.e., the capacitance is not constant in the accumulation or depletion regions of the *C–V* curve. The present status of correlation between electrical and physical measurements are discussed below.

Some production FABs monitor oxide thickness on both uniformly doped substrates and on PMOS and N-channel metal-oxide semiconductor (NMOS) test structures. The growth rate of SiO_2 is dependent on doping type (*p* vs. *n*) and concentration especially for lightly doped regions. Therefore, the gate oxide (here we refer to >2 nm SiO_2) must be measured on test structures that have implants representative of the channel region of a transistor. This means that when tight control of oxide thickness is required the oxide thickness in both the channel in the PMOS and in the NMOS regions must be measured. The assumptions made in the evaluation of the perfect MOS capacitor, such as uniform doping, are no longer valid. The interface between the silicon and the "bulk like" dielectric contributes to the measured capacitance. When this non-bulk oxide capacitance is constant or linearly varies with oxide thickness, correlation with optical (ellipsometric) measurement is possible.

Depletion of charge in the poly-silicon gate and quantum states in the bent bands at the interface of the crystalline silicon with the gate dielectric layer alter the *C–V* behavior described above. In Figure 24.31, we show the *C–V* data for thermally grown SiO_2 ranging in thickness from 2.5 to 1.5 nm [61]. The 2.5 nm oxide has classical *C–V* behavior while 1.5 and 2 nm oxides show the effect of quantum behavior and poly depletion. A procedure for removing both quantum and depletion effects from *C–V* data has been described [61,63]. The resulting oxide thickness can be directly compared to ellipsometric measurements. When the oxide is very thin, the voltage applied to the poly-silicon gate drops inside

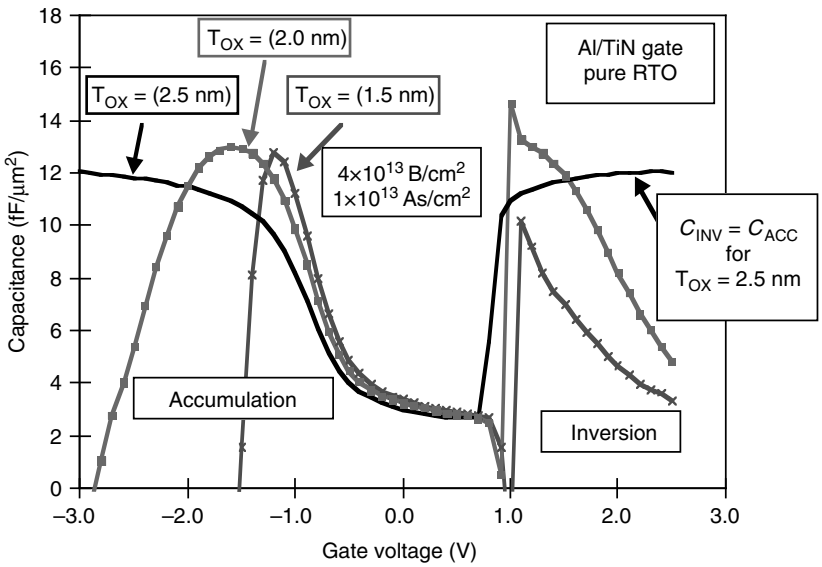


FIGURE 24.31 Capacitance–voltage data for thin oxides using Al/TiN gate electrodes. All data is courtesy George Brown, Texas Instruments. *C–V* data for 2.5, 2.0, and 1.5 nm oxides made using rapid thermal oxidation show the increased non-ideal capacitance in accumulation and depletion as the oxide becomes thinner.

the poly-silicon gate electrode. This effect is a function of oxide thickness and poly silicon doping. Modeling of the "poly depletion effect" has shown that for a 3.5 nm oxide the ratio of measured gate capacitance to true oxide capacitance is about 0.84 for a poly gate doping level of 2×10^{20} and $\sim$0.75 for 5×10^{19} [62]. For a 1.5 nm gate oxide the capacitance ratios are 0.7 and $\sim$0.55 for these gate doping levels [62]. Band bending at the silicon substrate–gate oxide interface allows the formation of quantum levels. When a negative voltage is applied and electrons are drawn towards the silicon substrate–gate dielectric interface, the accumulated electrons fill these quantum levels. The filled quantum levels increase the amount of band bending, and cause the center of the accumulated charge to shift away from the interface. These effects change capacitance significantly and must be accounted for during electrical measurement of oxide thickness. In Figure 24.32 both corrected and uncorrected theoretical *C–V* plots for a sub 3 nm SiO_2 gate oxide are shown.

Measurement of very thin SiO_2 with most test equipment requires the use of high frequency ($\sim$1 MHz) and transistor structures that have a large width to gate length ratio [62,63]. Direct tunneling through thin oxides distorts the *C–V* measurement which can be avoided by use of high frequency (non-static) *C–V*. Static *C–V* measurement have been extended to $<\sim$2.5 nm SiO_2 by use of special leakage compensation circuitry and numerical correction [63]. *C–V* metrology should be used for gate thickness on uniformly doped substrates. Simulation of *C–V* data for channel doped structures is not available at this time.

Recently, a new method of obtaining non-contact, *C–V* like data called the "Quantox" has been introduced. This tool can provide oxide thickness, flat band voltage, and carrier lifetime data [64]. A corona charge, *Q*, makes the top "gate" for the capacitance measurements, and provides the bias sweep. The Kelvin probe measures the surface voltage at each *Q* bias point, and the surface photovoltage (SPV) is the transient surface voltage measured at each *Q* bias point when a light is flashed to photoflatten the bands. This results in a low frequency non-contact *Q–V–SPV* curve. This is shown in Figure 24.33a and b. The author notes that there are differences between traditional *C–V* and Quantox measurements. The rate of voltage sweep is different from traditional *C–V* measurements, and the amount of tunneling of current during *Q–V–SPV* measurement is considerably less. By pulsing the corona charge to deep deplete

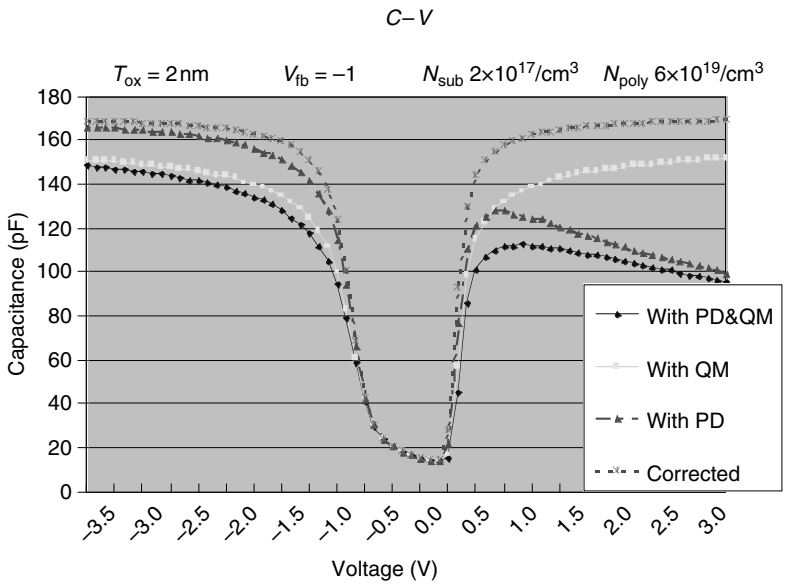


FIGURE 24.32 Theoretical *C–V* simulation showing quantum and depletion effects.

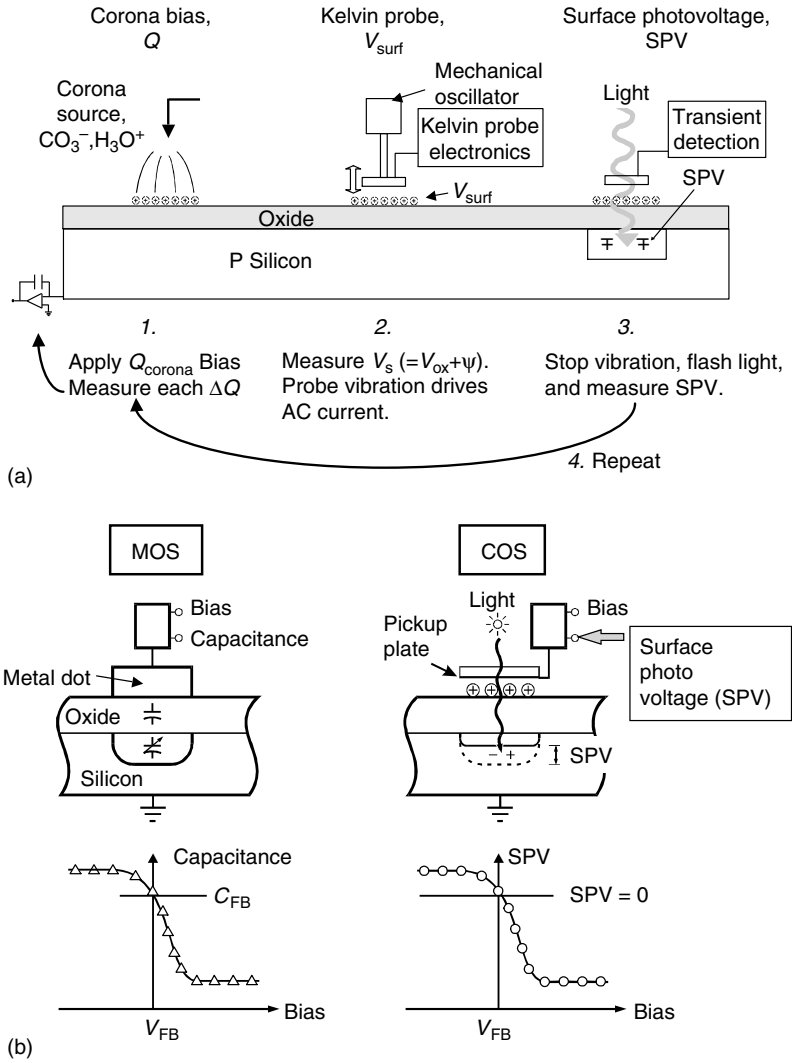


FIGURE 24.33 (a) Quantox measurement of surface photovoltage (SPV) vs. bias voltage. This figure shows the three components of a Quantox system. A corona charge is used as the gate for a corona-oxide-semiconductor (COS) capacitor measurement. The bias is swept by applying different corona charges. The Kelvin probe measures the transient SPV at each corona (Q) bias point after the bands are flattened by the flashing a light. Figure courtesy Steve Weinzierl, Keithley Instruments. (b) Comparison of traditional capacitance–voltage with quantox *Q–V–SPV* measurements. Figure courtesy Steve Weinzierl, Keithley instruments.

the silicon from inversion, non-contact measurements of carrier density and generation lifetime to a controlled depth are also provided. The comparison of Quantox data (not corrected for quantum and depletion effects) with ellipsometric data shows a strong correlation. Existing data shows that repeatability is 0.03 nm, 1 σ. Some workers believe that Quantox measurements are not effected by hydrocarbon buildup on the oxide surface. In addition, poly depletion corrections are not required.

Current–Voltage measurements also show a strong correlation to oxide thickness and may have better resolution to changes in oxide thickness than ellipsometry [60, 61, 63, 64]. Typical current–voltage data is shown for oxide thicknesses between $\sim$3 and $\sim$1.5 nm in Figure 24.34. Since the *I–V* data is a strong function of oxide thickness (10 Å/cm for a 0.04 nm decrease in oxide thickness), measurement resolution

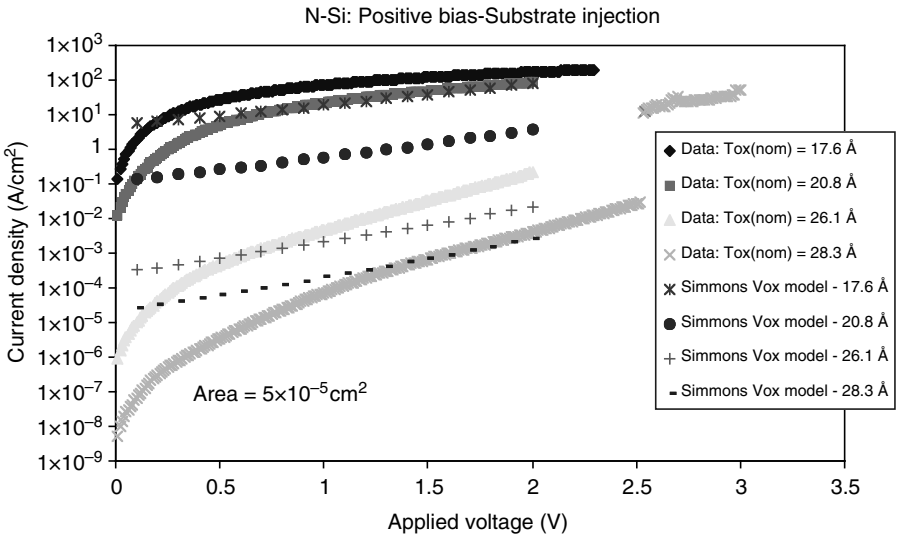

N-Si: Positive bias-Substrate injection
Current density (A/cm²)
Applied voltage (V)
Area = 5×10⁻⁵cm²
Data: Tox(nom) = 17.6 Å
Data: Tox(nom) = 20.8 Å
Data: Tox(nom) = 26.1 Å
Data: Tox(nom) = 28.3 Å
Simmons Vox model - 17.6 Å
Simmons Vox model - 20.8 Å
Simmons Vox model - 26.1 Å
Simmons Vox model - 28.3 Å

FIGURE 24.34 Current–voltage data applied to oxide thickness measurement. At 1 V bias, the current varies over seven orders of magnitude as the oxide thickness changes from ~3 to ~1.8 nm. Simulated data is also shown. Figure and data courtesy George Brown, Texas Instruments.

appears better than the approximate 0.25° change in Δ per angstom estimated for ellipsometry at 632.8 nm. On the basis of the measurement procedure reported by Brown and co-workers, the thickness associated with measured current must be calibrated using another measurement. The *I–V* characteristics should depend on doping characteristics such as channel dose. Unless calibrated to *C–V*, *I–V* thickness does not depend on knowledge of the static dielectric constant.

24.3.3 New Methods of Measuring Gate Dielectric Thickness and Nitrogen Concentration

X-ray photoelectron spectroscopy is a well known materials characterization method that determines the elemental composition along with their chemical state [65]. The information depth for XPS is roughly the top 5 nm or less. Depth dependent information can be determined using either ion sputtering or by measuring the XPS signal at different angles of x-ray incidence [66]. Recently, a clean room compatible XPS system capable of rapid throughput has been introduced [67]. The first data for the precision of film thickness and nitrogen concentration are promising and worth of mention. This method can also be applied to high κ composition and thickness metrology. The spot size of XPS is determined by the size of the area irradiated by the x-rays and the region of this area that generates electrons that enter the detector. This area is 9 μm in diameter which is larger than that analyzed by optical methods.

Electron microprobe analysis is another materials characterization method that has been adapted for clean room based metrology. In this technique, x-rays are emitted after irradiation by an electron beam. In one commercial version, low energy x-rays from elements such as nitrogen and oxygen are measured using wavelength dispersive spectrometers. In order to generate enough x-rays to have favorable signal to noise, a relatively large diameter electron beam with a relatively large total current [68] is used. Preliminary evaluation indicates that the precision for nitrogen concentration is ~1% of the dose between 0.8 and 5% nitrogen concentration and the precision for thickness is ~1% of the thickness for

thin films [65]. The analysis area is between 10 and 100 μm in diameter, and the electron beam energy is from 0.2 to 10 keV. This method can also be applied to dopant dose metrology.

24.3.4 Doping Process Control

The predominant method of dopant processing is implantation of ionized dopants. For the purposes of this chapter, we will refer to three types of implant steps: high energy—medium dose [10^{13} P^+/cm^2 at 700 keV, 10^{13} B^+/cm^2 at 1.5 MeV] to form retrograde well structures and [B^+ and P^+10^{12}–10^{13} ions/cm^2 at 100–200 keV] for N and P wells; medium energy—high dose [10^{15} As^+/cm^2 at 10–80 keV; 10^{15} B^+/cm^2 at 0.5–40 keV or 10^{15} BF_2^+/cm^2 at 5–200 keV] to form the source and drain and [10^{13}–10^{14} As^+, P^+, B^+, or BF_2^+/cm^2 at 0.2–100 keV] to form the lightly doped drain (also called drain extension); and low energy–low dose [10^{11}–10^{12} As^+, P^+, B^+, or BF_2^+/cm^2 at 0.2–100 keV] for the threshold voltage implant. Each of these types has a different process tolerance and thus requires a different precision from the metrology tool [69]. Another issue is uniformity across the wafer. Several references have discussed implant process control including SPC and uniformity [70,71].

The three most prevalent methods of dopant dose control are four point probe, optically modulated optical reflection (commercially known as the Thermawave), and secondary ion mass spectrometry (SIMS) depth profiling. Four pt probe is used for high dose plants and Thermawave for medium to low dose process control [69]. SIMS can be applied to junction and dose measurements for all doses. The precision of optical densitometry has recently been improved, and new methods have been proposed. In this section, we discuss all these methods.

24.3.4.1 Four Point Probe

The four point probe method characterizes the active dopant dose by measuring the resistivity and relating this to the active carrier concentration. Two reviews of the fundamentals of four point probe measurements are available [72,73]. The four point probe method of measuring resistivity is based on the use of a linear array of four probes in which two probes carry the current and two probes measure voltage. The resistance of a two probe configuration includes contributions from probe contact and spreading resistance [72]. The two additional probes are used to sense voltage due to the current passing through the sample between the other probes, and current through these probes is minimized and, thus, the spreading resistance and contact terms are minimized. An alternative method involves analysis of the resistance and currents when different pairs of the four point probes are used for injecting current and measuring potential. The effect of the spreading and contact resistance can be mathematically eliminated, and thus this is a second approach to removing these effects. Therefore, the contact and spreading resistance terms cancel out in the four point configuration when the voltage probes are used instead of the two probe approach [72]. Typically, the probes are equally spaced and the current runs through the outer probes [72]. The probe spacing on traditional systems is between 0.5 and 1.5 mm [72], and greatly improved probe tips (larger radius probe tips for improved precision) are required for 180 nm technology generations. The measured resistivity is a function of the sample shape. The equation for the resistivity ρ is: $\rho = 2\pi sF(V/I)$, where s is the probe spacing, V is the voltage at the voltage sensing probes, and I is the current through the current carrying probes. F is the correction factor that accounts for sample shape, and it is a function of thickness, lateral dimensions, and wafer edge proximity.

Four point probe measurements are done using monitor wafers having special characteristics. After implantion, the monitor wafers must be annealed to restore crystallinity and activate the dopant [69,70]. These monitor wafers must have resistivity characteristics that permit measurement of the implant dose of interest. For example, it would be difficult to measure the change in resisitivity of a low resistivity wafer (e.g., p+wafer) due to a low to medium dose implant. Epi wafers such as "p" epi on p+substrate also create problems since the current can pass through the lower resistance substrate which is away from the implant. Implant doses $>10^{14}$ ion/cm^2 are easily measured on wafers having a resistivity $\geq 20\ \Omega$-cm. Doses in the 10^{11}–10^{12} ions/cm^2 range may be measured using high wafers with a high resistivity when

the correct annealing and surface conditions are used. Yarling and Current suggest either "p" or "n" type wafers with a resistivity $>100\ \Omega$-cm for low dose samples [70].

Correct annealing conditions are critical. Due to the processing characteristics of different anneal/dopant combinations, the anneal must be optimized for the parameter being tested. Implant dose alone is often tested by long, high temperature furnace anneals, while implant energy is best tested with a short rapid thermal processing (RTP) anneal or, better yet, with a blocking layer (e.g., oxide layer) which only tests the deeper part of the implant. Similarly, RTP temperature control for implant diffusion is most optimally tested using an implant near the solid solubility. The doping level is most sensitive to the temperature under these conditions.

24.3.4.2 Optically Modulated Optical Reflection

Optically modulated optical reflection is the monitoring of thermal wave propagation in an implanted wafer. The commercial system used for this measurement is often called by the supplier's name, Thermawave. Thermal waves are produced by an argon ion laser modulated at a frequency of 0.1 to 10 MHz [74]. These waves heat the adjacent surface of the wafer changing the volume of the silicon near the surface. The volume change alters the optical properties of the surface, and this is probed by measuring the reflectivity change using light from a second laser. The amount of heat transfer and the volume change per unit heat depend on the implant dose and implanted species because heat transfer is effected by the implant process damages the silicon lattice structure while adding dopant atoms to the lattice. The sample spot size is approximately 1 μm, and measurements of doses between 10^{11} and 10^{15} atoms/cm^2 have been reported. Optically modulated optical reflection is distinguished by its sensitivity to low dose implants. Each implant ion typically produces 100 to 1000 damage sites so the effect from each ion is multiplied. After a certain dose is reached, there is little change in implant damage, and sensitivity to high dose is hampered.

24.3.4.3 Secondary Ion Mass Spectrometry

Secondary ion mass spectrometry can provide dopant dose, and since SIMS data is in the form of a depth profile, it naturally lends itself to junction measurements. The first dynamic SIMS systems equipped with product wafer capable stages were delivered in 1997. Since SIMS has been described in the chapter on Materials Characterization, only the application to dopant process control is discussed here [75]. SIMS is capable of simultaneously monitoring multiple implant species in the same test structure. Test structures are typically 100 μm$\times$100 μm and larger and have been routinely incorporated in the scribe lines of product wafers. Test structures are also routinely placed in the die of product wafers especially during process and pilot line development.

Implant dose is measured by integrating the secondary ion signal of the dopant element from the surface to the depth of the implant. Shallow implants remain a significant challenge for SIMS characterization [75]. The secondary ion signal is strong function of the matrix that the implanted ion resides in (oxide, interface, or bulk silicon). In addition, quantifiable secondary ion signals require that the ion sputtering process reach a steady state. When using energetic primary ion beams such as 5 to 10 keV O_2^+ or 5 to 10 keV Cs^+, the steady state is not reached until a considerable depth (sometimes up to 50 nm) is sputtered away. Recent studies show that for 300 eV O_2^+ primary beam, a steady state is reached for crystalline, oxidized crystalline, or amorphized silicon by the time 0.7 nm has been sputtered away [76,77]. This has driven the use of low energy ion beams for shallow junctions [76–78]. Stable, very low energy ion beams around 100 eV were introduced in 1997. A rule of thumb for depth profiling is that the ion beam energy should be half the implant energy. The optimum conditions for B implant measurement depend on whether or not oxygen gas is used to dose the sample surface along with a primary ion beam of O_2^+. When possible, very low energies (down to 200 eV or less) are used for the primary beam. This can slow down the analysis procedure if the ion beam current is low, and beam energy and current should be selected according to conditions that depend on the lowest concentration of B that one needs to determine [76–78]. The sputter yield decreases below 1 keV primary beam energy and the detection limit is approximately 10^{17} atoms boron/cm^3 for 100 eV O_2^+ at normal incidence. When oxygen dosing is used,

low energy primary ions at incidence angles from 45 to 55° have shown satisfactory results. When no oxygen dosing is used, then normal incidence should be used for the primary ion beam. Shallow arsenic implants can be analyzed using Cs^+ at 60° incidence and 500 eV energy.

24.3.4.4 Junction Depth Measurement via Carrier Illumination

Carrier Illumination is a new, non-destructive method of measuring junction depth. Borden has described Carrier Illumination in depth [79]. The junction was considered to be the depth at which carrier concentration drops to $10^{18}/cm^3$. Due to the increase in channel doping, a value of 1×10^{19} is becoming more relevant. The method first excites excess carriers using a 2 μm spot laser. The carriers form a quasi-static distribution that piles-up at the edge of the doped layer. The excess carriers are due to the change in the direction of flow of the excess carriers from vertical in the doped layer to a radial horizontal flow in the substrate. The depth of this junction is determined by measuring the reflectivity of a second laser [7,80].

24.3.5 Metrology for Measurement of Stress Enhanced Carrier Mobility

Stress metrology faces the difficult challenge of controlling processes used to increase carrier mobility through stress. As discussed above, greater carrier mobility increases transistor drive current and thus transistor switching speed [81]. Strained silicon substrates are also being investigated for this purpose [82]. First, the diverse set of processes is used to increase mobility as discussed, and then the metrology methods and challenges are reviewed.

Stress is the force applied to a material, and it can be compressive or tensil. Stress can also be uni-axial, that is along one crystallographic direction, or bi-axial or along two, perpendicular crystallographic directions. Process induced stress is typically uni-axial, while the lattice mismatch with the substrate induces a biaxially strained silicon surface layer. Stress induced changes in physical properties are typically modeled by as if the silicon is a continuous elastic material. Recent publications indicate that uni-axial stress has some significant advantages over strained silicon layers [83]. The amount of stress required for improving carrier mobility is less for uni-axial stress. In addition, the process based approach allows one to use compressive stress for p-channel devices and tensile stress for n-channel devices while the substrate approach results in only tensile stress [83].

Process based increase in carrier mobility has been used in the manufacture of the 90 and 65 nm technology nodes. Processes induce nearly uni-axial, tensile stress to increase electron mobility and uni-axial, compressive stress to increase hole mobility. Intel uses tensile silicon nitride layers above the NMOS and replaces the source and drain in the PMOS with silicon germanium to compressively stress the channel [84]. Although direct measurement of the stress in the buried channel is impossible, one can surmise how process control is done. Measurement of silicon nitride thickness can provide control of a stable silicon nitride deposition process for the NMOS. It seems likely that a combination of CD and Ge concentration measurement could be used for controlling the PMOS mobility. Recent publications discuss the modeling how the Intel Process induces compressive stress along the ⟨110⟩ direction of silicon increases hole mobility [85]. Texas Instruments has found that use of recessed SiGe source and drain extensions improves PMOS drive current by 35% [86]. Fujitsu has shown that by changing process conditions, the silicon nitride layer can be used to impart both tensile and compressive stress [87]. IBM uses the stress induced by the shallow trench isolation to stress PMOS channels [88]. The amount of stress can be altered by changing the distance between the STI and the edge of the gate electrode. Future approaches to stress induced improvement of carrier mobility include the use of metal gates and strained silicon substrates [89].

Strained silicon substrates are grown on top of silicon germanium layers where the lattice mismatch provides the stress. Strained silicon can either be left on top of the SiGe or transferred to a wafer with a surface oxide layer producing the so-called sSOI.

Strain can be measured by a number of different methods such as Raman spectroscopy, x-ray Diffraction, and Photoreflectance. The average stress across a wafer can be measured using the change in wafer curvature. Some of these methods are briefly described below.

Raman spectroscopy measures strain using the shift in the wavelength of the silicon optical phonons. The shift can be related to the stress through the elastic equations. In the future, nano Raman systems based on near field optical microscopes will push the spatial resolution below 200 nm. High resolution–x-ray diffraction measures small changes in lattice constant (strain). Measurement of patterned wafers requires large test areas. Photoreflectance Spectroscopy can measure strain in un-patterned wafers. The electronic transition between energy levels occurs at specific energies that change when the lattice structure is stretched or compressed. Bi-axial stress splits energy levels that degenerate in un-strained silicon or germanium. In particular, the E1 transition energy at 3.392 eV is used to monitor strain.

The average stress across a wafer can be calculated from the change in wafer curvature after film deposition. Die level stress can be determined from local wafer curvature using Interferometry. The new Coherent Gradient System uses a referenced interferometer to measure local curvature of the wafer. Using well known algorithms, the local stress changes can be calculated from the local wafer curvature changes when a wafer is measured before and after a process step.

24.4 Interconnect Process Control

In this section, we will refer to interconnect processes, as those that begin with the contact between the transistor and the first level of on chip interconnect metal (Metal 1) and end with the passivation layer over the final on chip metal level. In Figure 24.35, we illustrate the typical process steps that are controlled by metrology. The process schemes for interconnect are expected to be a mixture of traditional etched metal/inter level dielectric and Damascene (also known as inlaid metal) processing. Chemical Mechanical Polishing and copper will be used in both types of interconnect processes. The interconnect via and contact (also known as plug) material could be tungsten or aluminum with titanium nitride barrier layers or copper with barrier (possibly Ta) in the future. Routine physical metrology needs for interconnect processes include metal and dielectric film thickness, step coverage, CMP end point and flatness, and particle/defect control. Particle detection and control is covered in the chapter on Contamination Free Manufacturing. When interconnect processes are stable, film stress is not routinely monitored. The CMP processing may replace the need for measurement of B and P in reflow glass inter-layer dielectric (ILD) films. Electrical metrology needs include testing of contact/via resistivity and metal level defects.

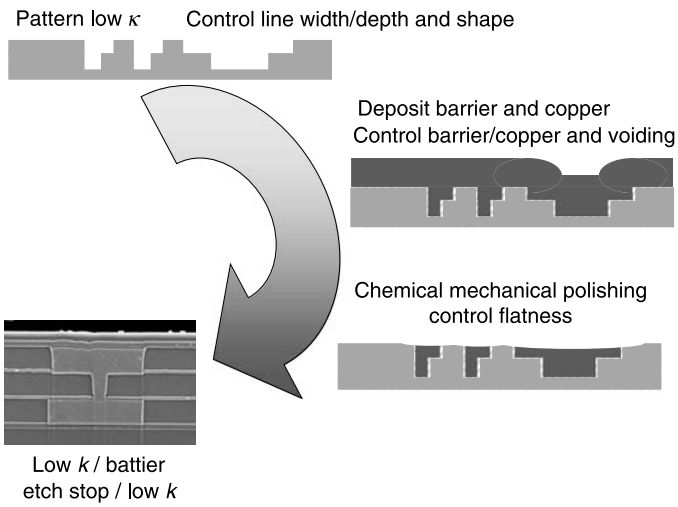


FIGURE 24.35 High level overview of interconnect metrology.

Process control needs also include the need to measure voids in copper lines and large "killer" pores in porous low k materials. Although porous low k is (at the time of writing) a research material, great progress has been made in the measurement of pore size distribution.

24.4.1 Interconnect Film Thickness

In this section, we discuss film thickness measurement for contacts (e.g., $TiSi_2$), barrier layer (e.g., TiN, TiW, Ti/TiN stacks, or $CoSi_2$), metallic anti-reflective coatings, interconnect dielectric layers, and all metal interconnect films (e.g., Al or Cu). Ideally, film thickness metrology would be done on patterned wafers and include measurement of the liner film thickness inside a via/contact structure. Most film thickness measurements require large, flat sampling areas. A number of commercially available methods have been applied to metal and barrier layer film thickness measurement such as acoustic [90,91], resistivity by Four Point Probe [92], x-ray reflectivity [93], XRF [94], and a new method known as Metal Illumination (MI)™ [95]. A newly introduced, in-line method that is very similar to Electron Microprobe uses a relatively large electron beam to excite XRF. This is covered in the section below on determination of metal film thickness using x-ray Methods. Optical reflectivity and ellipsometry are both used for dielectric film thickness [96]. It is useful to note that ellipsometry can measure the thickness of very thin barrier layers. Off-line measurement of patterned metal thickness is possible using scanning electron microscopy and energy dispersive x-ray spectroscopy (EDS) if careful modeling of x-ray emission is done for the film stack and structure that one is analyzing.

24.4.1.1 Interconnect Metal Film Thickness Measurement Using X-Ray Methods

Metal and barrier layer film thickness can be measured in-line by commercially available x-ray reflectivity and XRF. Both methods can determine barrier thickness for barrier films under the seed copper layers. Measurement of barrier layers under the thicker electrochemically deposited copper layers used in the fabrication of metal lines is more difficult, but reports of successful use of XRF exist. Measurement on patterned layers is possible using the electron beam excited, XRF approach embodied in the Matrix 100™ [97]. Although x-ray reflectivity may be applied to patterned films, its use is not widely reported at this time. Innovations in x-ray optics make it difficult to know what the true limits of x-ray methods concerning patterned layer measurements.

X-ray reflectivity is a very powerful method of characterizing film thickness. x-ray reflectivity is also sometimes referred to as grazing incidence x-ray reflectivity (GI-XRR). In GI-XRR, the angle of incidence of a well collimated, monochromatic x-ray beam is reflected off a flat sample over a range of incident angles [93]. The intensity of the specular reflection is used for film thickness determination. Non-specular x-ray scattering can also be measured, and film and interface roughness analyzed. Interference patterns in the form of intensity oscillations are formed when GI-XRR is done on single and multi-layer thin film samples. Reflectivity from a homogeneous substrate, a single, and a two layer film is shown in Figure 24.36. The film thickness of a single layer film can be determined from the angular difference between the peaks of subsequent intensity oscillations. Two periods of intensity oscillation are present when analyzing a two layer film. Because the wavelength of a monochromatic x-ray is accurately known, thickness can be accurately determined. Density measurement is more difficult.

It is important to note that in-line measurement of x-ray reflectivity is done using an optical path that simultaneously collects reflected x-rays at multiple angles. Specially designed optical paths and detectors allow rapid measurements making it possible to measure multiple wafers per hour inside a clean room. In some sense, x-ray reflectivity and ellipsometry are similar in that both require development of appropriate models to interpret the data. The key to successful modeling is incorporating interfacial characteristics that allow for high precision and appropriate "Goodness of Fit" to the data. Due to the difference in electron density, very thin (0.5 nm) barrier layers have been measured below seed copper films. Due to x-ray absorption, measurement of barrier layers under electroplated copper is usually not considered to be possible.

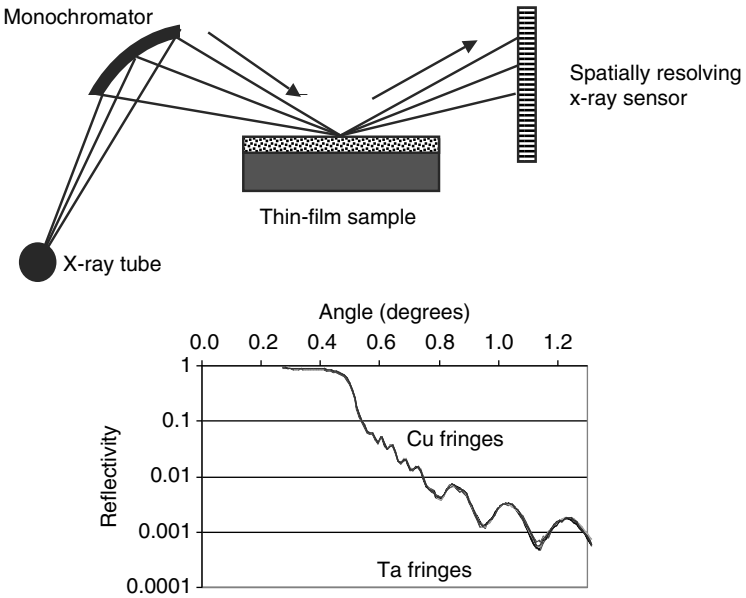


FIGURE 24.36 Diagram of an x-ray fluorescence film thickness tool. The Kevex tool uses collimated probe beam and collimators for the fluoresced x-rays. Figure courtesy David Wherry and Ed Terrell, Kevex.

In-line measurement of metal film thickness has been done using both energy dispersive and wavelength dispersive XRF [94]. Micro-focus tools have a spot size of 50 μm while non-micro-focus systems typically have millimeter size spots. Film thickness can be quantified from 1 nm to 10 μm. Repeatability, which is the short term contribution to precision, is <0.5 nm for 50 nm Ti and TiN films and <1 nm for 1 μm Al films. Using modeling, film stacks can be characterized. This method applies to any barrier layer, TiN type antireflective coating, and both Al and Cu interconnect metal films.

Typically, XRF is used to measure the thickness of thin films that are composed of an element or elements different than the silicon or silicon dioxide layer below, e.g., Al or TiN. In addition, the substrate composition must be constant. This greatly simplifies calculation of film thickness from the fluorescent intensity of the element(s) that compose the film. X-ray fluorescence is created by both the incident x-ray beam and fluorescence from surrounding material such as the substrate below the film. The thickness, t, of a film of element (i), I_i, is related to the measured fluorescent intensity of an infinitely thick film of element (I), $I_{\propto i}$, the film density, ρ, and the effective mass attenuation coefficient, μ_i^*, by [94]:

$$t = (1/\rho\mu_I^*)\ln(1-(I_i/I_{\propto i}))$$

The quantity μ_i^* is a function of x-ray wavelength, angle of incidence ψ', and angle of exit to the detector ψ'':

$$\mu_i^* = \mu_{i\lambda}\,\csc\psi' + \mu_{i\lambda}\,\csc\psi''$$

$\mu_{i\lambda}$ is the mass attenuation coefficient of element I at the wavelength of x-rays used to excite the sample. Daily calibration is required due to the small, day to day variations in x-ray tube intensity and detector efficiency.

In Figure 24.37, the x-ray optical paths for one type of XRF systems are shown. The x-ray beam is used as a means of exciting XRF from the film. Since the films are poly-crystalline, x-ray diffraction is difficult to minimize unless the film is highly textured. Although the depth of penetration depends on the angle of incidence and the adsorption of the material, it is of the order of microns. X-ray fluorescence can also be observed from more than a micron below the sample surface [94]. Analysis of XRF is made difficult by

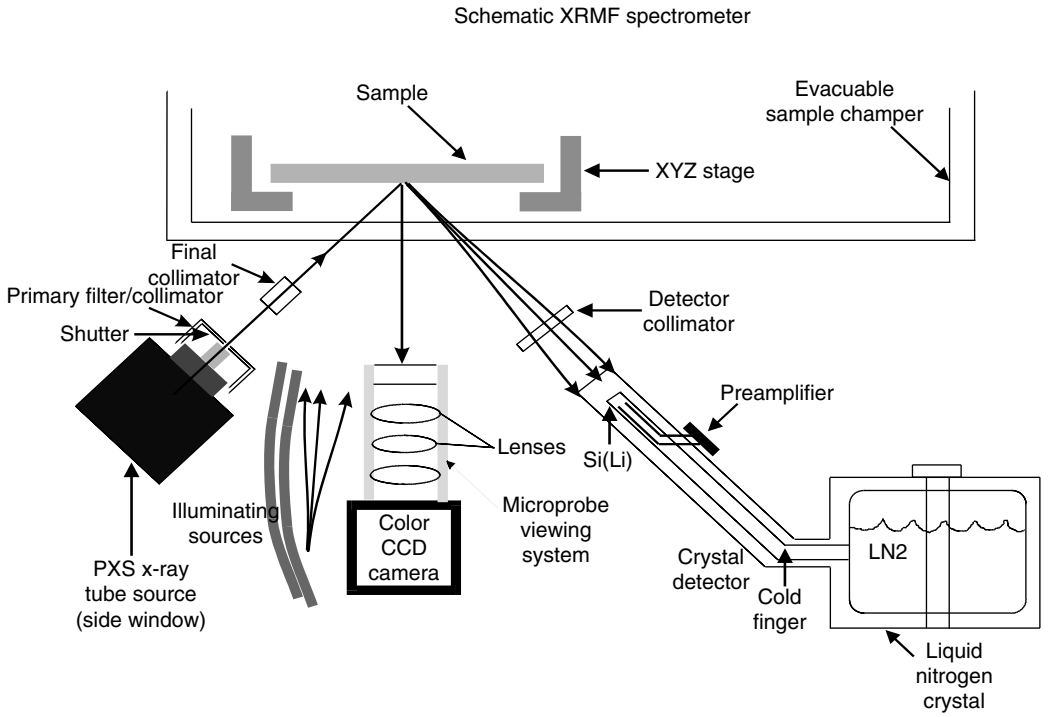


FIGURE 24.37 X-ray reflectivity based film thickness measurement In-line x-ray reflectivity throughput is increased by the use of optics that allow the entire reflectivity vs. angle data to be obtained at one time. X-ray reflectivity data for a copper film on a tantalum barrier layer is shown.

the adsorption of fluoresced x-rays and subsequent re-emission. Therefore, reference materials are critical to the success of this method. In addition, x-ray tube emission and x-ray detectors require daily recalibration.

24.4.1.2 Metal and Interconnect Dielectric Film Thickness Measurement Using Acoustic Methods

In this subsection both of these non-destructive methods for measuring metal film thickness are described. Considering the original application of one acoustic method, dielectric film thickness measurement may also be possible.

One of these systems is based on impulsive stimulated thermal scattering (ISTS) which is now referred to as laser induced surface acoustic waves. Commercially, the method is now known as SurfaceWave. Detailed descriptions of this method are available in the literature [90]. The acoustic wave travels parallel to the surface of the sample in ISTS, in contrast to picosecond laser ultrasonic sonar where the acoustic wave travels downward. The ISTS is a rapid (<1 s. per data point), small spot (15 μm$\times$30 μm) film thickness method which has outstanding precision when applied to single layer films. The optical path has been designed for a long working distance and with solid state, long life lasers making it suitable for in situ sensor applications.

In ISTS, a pair of optical pulses each having a duration of a few hundred picoseconds are overlapped in time and space on a sample's surface. Optical interference between the crossed excitation pulses forms a spatially varying interference or "grating" pattern of alternating light (constructive interference) and dark (destructive interference) regions. Formation of the grating pattern is illustrated schematically in the inset of Figure 24.38a. The grating fringe spacing Δ (or wave number q) is given

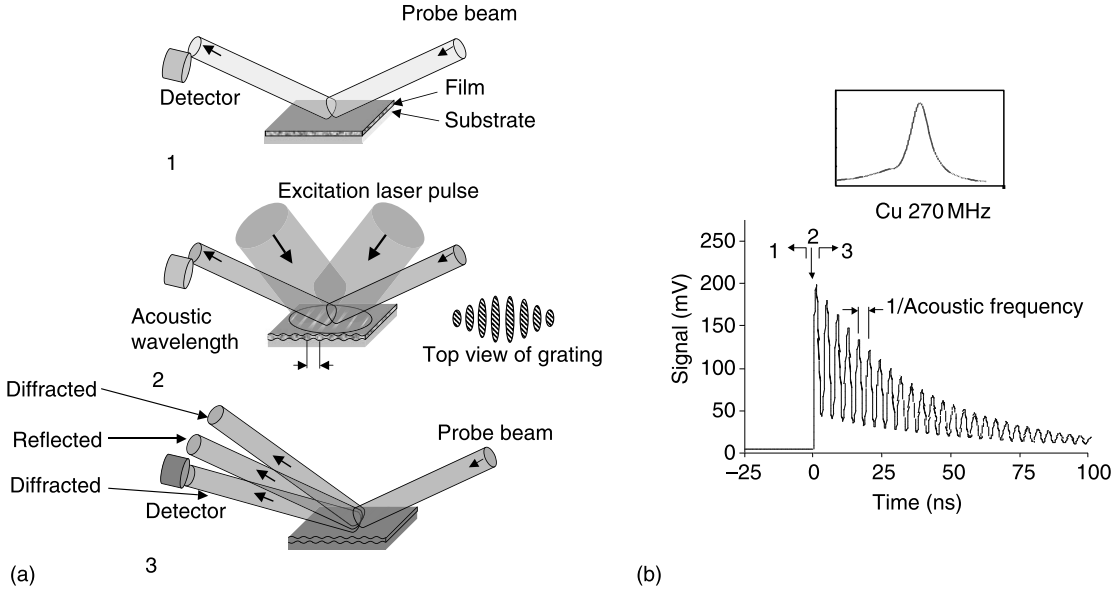


FIGURE 24.38 (a) The time dependent modulation of reflectance due to the acoustic wave produced by impulsive stimulated thermal scattering (ISTS; Surface Wave) from a single copper layer. The three steps in a measurement are: 1. The probe beam strikes the surface, 2. The grating structure is formed from the overlapping parts of the excitation laser pulse and the acoustic wave is formed, 3. The acoustic wave travels away from the excitation area and the traveling surface ripples diffract the probe beam in and out of the detector. Figure courtesy Michael Gostein, Philips Advanced Metrology Systems. (b) Acoustic frequency of Cu film.

by: $q = 4\pi \sin(\Theta/2)/\lambda = 2\pi/\Delta$. The crossing angle Θ of the two excitation pulses. The sample absorbs radiation in the light regions, resulting in a mild heating and thermal expansion that launches coherent acoustic waves whose wavelength and direction match those of the interference pattern with wave vector $\pm q$ [90]. The acoustic waves generate a time-dependent "ripple" on the sample's surface. The depth of modulation oscillates at the acoustic frequency, which is determined by the sample's mechanical (e.g., elastic) and physical (e.g., thickness) properties and by the boundary conditions (e.g., adhesion) between the different layers in the sample. At the time that the acoustic waves are propagating outward from the excitation site, heat flows from the heated grating peaks to the unheated grating nulls at a rate determined by the sample's thermal diffusivity. One advantage of this acoustic method is that by changing the angle Θ, the wave vector is changed and properties such as sample density can be experimentally determined. Metal film thickness methods, such as XRF Rutherford backscattering, and acoustic, all require knowledge of the sample density. The ISTS can provide experimental density information specific to the metal film fabrication process. Changes in film density can also be monitored.

The acoustic response is measured in its entirety by diffracting a probe laser pulse having duration of several hundred microseconds off the surface ripple to form a pair of signal beams (the $+1$ and -1 diffracted orders). One of the signal beams is detected to generate a light-induced signal waveform [90]. This diffraction mechanism is indicated schematically in the inset of Figure 24.38a. This figure also shows time-dependent data taken from a copper film. These data were collected during a period of one or two seconds. During its first 70 ns, the signal waveform oscillates according to the copper's acoustic frequency and decays primarily due to the travel of the acoustic waves away from the excitation cite into the rest of the sample. Figure 24.38b shows the power spectrum of the copper data in Figure 24.38b to be centered at about 270 MHz. To determine film thickness, the frequency is analyzed as described in the above-mentioned references [90].

Figure 24.39 illustrates the accuracy of the ISTS measurements, showing, respectively, the results from a set of copper/oxide/silicon and tantalum/oxide/silicon samples having thicknesses ranging from 20 to 180 nm. These data show the correlation between the center-point film thickness determined using ISTS (y axis) and a conventional four point probe (x axis). Film thicknesses were calculated from the four point probe data using resistivity values that are within 2% of the bulk resistivities for both tantalum and copper. The line in each of the plots has a slope of one and is used as a guide to the eye. For both the

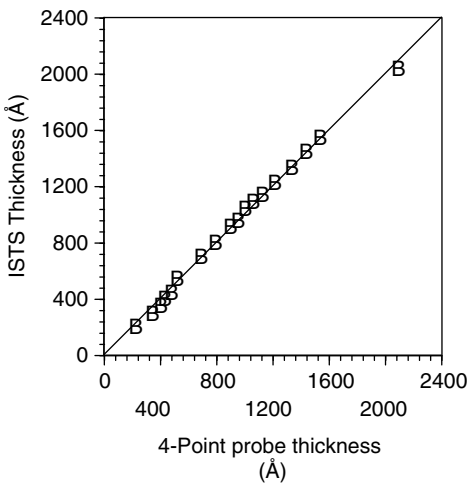


FIGURE 24.39 Comparison of ISTS and 4-point probe measurement of film thickness. The data was taken using the InSite 300, and the figure was provided by John Hanselman of Active Impulse Systems. This technology is now provided by Philips AMS.

tantalum and copper samples, the correlation between the ISTS and four point probe thickness measurements is excellent over the entire sample set. On average, these values are within 0.6 and 2.9% of each other, respectively. Deviation between the two measurements may be due to a variety of factors, such as the film's resistivity depending on film thickness of grain structure. Such dependence seems particularly evident in the tantalum data, as the deviation of the data from the solid line systematically increases with film thickness.

Impulsive stimulated thermal scattering can measure one or more layers in a multilayer structure. Two different approaches are used based on the thickness of the films [90]. The first method is applicable when both layers are thick (e.g., $> \sim 1000$ Å) and have different acoustic properties. Two acoustic wavelengths are recorded, one short and one long. The two wavelengths probe the sample to different depths. The short wavelength response is relatively more sensitive to the top layer than the bottom one, as compared to the long wavelength response. By analyzing both responses and fitting with a first-principles model, both layer thicknesses can be determined [90]. The second method is applicable when there is a thinner layer underneath the top layer, and both layers have different thermal properties. This method is used for barrier–copper seed applications. In this case, only one acoustic wavelength is used. The frequency response of the signal waveform is sensitive roughly to the total metal mass, while the thermal decay time of the waveform is sensitive to the ratio of the two metals, due to their thermal contrast. Using this information, both metal thicknesses are determined [90]. Presently, a buried layer of down to 5 nm can be measured and controlled.

One advantage of ISTS is that it can measure patterned wafer samples. In the case of a sample that consists of metal lines, the acoustic wave travels away from the excitation area along the array of metal lines. The acoustic response of the metal line/insulator array must now be modeled instead of a blanket film. The barrier layer in this type of sample is on both sides and below the metal line. Arrays of lines or vias are treated in the analysis as films with "effective" elastic constants that are a composite of those of the consistuent materials, i.e., metal and dielectric. In many cases the composite elastic constants can be estimated by simple averaging formulas. More complex structures, e.g., those with pitch similar to the acoustic wavelength, for which modeling is complicated, require a calibration to determine the effective elastic constants. In addition, the presence of vias below the metal lines complicates modeling. Due to these complications, the response vs. line thickness is usually calibrated, and routine measurement of film thickness for patterned structures is possible for metal lines, pads, and vias. The CMP process control is a typical application of patterned wafer measurements [90].

Another acoustic film thickness system was commercially introduced in 1997. One system is based on Picosecond Ultrasonic Laser Sonar Technology (PULSE). A detailed description of the physical processes involved making measurements of this type is available [91]. The PULSE technique has a high (a small numerical value) precision, sampling speed (2 to 4 s/pt), spatial resolution (less than a 10 μm diameter spot size), and it can be applied to single and multilayer metal deposition processes. In a PULSE measurement for a sample consisting of a metal film deposited on a substrate (e.g., Si) an acoustic wave is first generated by the thermal expansion caused by sub-picosecond laser light pulse absorbed at the surface of a the metal. The temperature increase (typically 5°C–10°C) is a function of sample depth which results in a depth dependent isotropic thermal stress [91] which gives rise to a sound wave which propagates normal to the sample surface into the bulk. This wave is partially reflected at the interface between the film and substrate [91]. For an acoustic wave, the reflection coefficient R_A depends on the acoustic impedances Z (Z = density × sound velocity) of the film and substrate materials, and may be evaluated from the relation $R_A = (Z_{sub} - Z_{film})/(Z_{sub} + Z_{film})$ [82]. When a reflected wave (or "echo") returns to the free surface after a time τ, it causes a small change in the sample optical reflectivity, ΔR. This change is monitored as a function of time by a second laser probe. Based on the sound velocities of the materials making up the sample (which for most materials are known from bulk measurements) and the echo time, the film thickness d_{film} may be evaluated from the simple relation $d_{film} = v_s \tau$.

As an example, Figure 24.40 shows a PULSE measurement obtained for a sample consisting of a copper film deposited on top of a thin Ta layer (less than 20 nm thick) with a substrate consisting of a thick tetraethoxysilane (TEOS) layer (about 600 nm). The sharp feature observed in a range of time less than

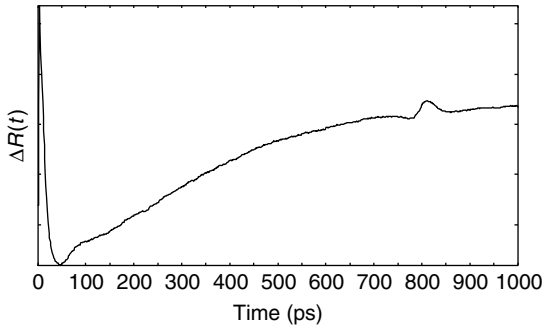


FIGURE 24.40 Acoustic measurement of copper film thickness using picosecond ultrasonic laser sonar (PULSE) technology. The time dependent reflectivity observed during PULSE measurement of a copper layer on 20 nm of tantalum on tetraethoxysilane (TEOS) is shown. The ultrasonic wave travels into the copper layer partially reflecting at the Cu/Ta interface and subsequently the Ta/TEOS interface. When the ultrasonic wave returns to the surface, it alters the optical constants of the surface and this is monitored by measuring the change in reflectivity. The transit time is directly related to the speed of sound and the layer thickness. The data taken using a MetaPULSe picosecond laser ultrasonic sonar system, and the figure is courtesy Rob Stoner, Rudolph Technologies.

about 50 ps is associated with relaxation of the electrons in the copper film which initially gain energy from the ultrashort light pulse. This energy is transferred to thermal phonons in the film which gives rise to an increase in its temperature by approximately one degree. The subsequent diffusion of heat out of the film and into the underlying TEOS occurs on a timescale of hundreds of picoseconds, and this is associated with the slowly decaying "background" signal observed in the figure. The sharp feature observed at a time of about 800 ps is the echo caused by sound which has reflected from the bottom of the copper layer and returned to the surface of the sample. To determine the film thickness from these data only the echo component is used; the background is discarded. From the product of the one way travel time for sound through the film (406 ps) and the sound velocity (51.7 Å/ps), the film thickness is found to be 2.10 μm.

The time-dependent reflectivity change $\Delta R(t)$ measured via PULSE depends on the physical mechanisms underlying the sound generation and propagation, and sound-induced changes in optical properties. These may be simulated with very high fidelity even for samples consisting of many layers with thicknesses ranging from less than 50 Å to several microns. Using an iterative procedure in which the thicknesses of layers in a model for a sample are adjusted until a best fit to the measured $\Delta R(t)$ is obtained; the commercial PULSE system obtains typical measurement precision for thickness of less than 0.1 nm (1 Å) with 90% confidence. This modeling methodology also allows other film properties such as density, adhesion, and surface roughness to be obtained along with thickness since these parameters have a predictable influence in the amplitudes and shapes of the optically detected echoes. In Figure 24.40, we show the PULSE signal from a 200 nm thick layer of TiN on top of aluminum. The ability to measure film density has considerable practical significance, especially for WSi_x, TiN, WN, and other materials whose composition and structure depend on many aspects of the deposition process such as pressure, target composition, temperature or gas mixture. The PULSE has been used to distinguish between TiN films differing in density (or composition) by only a few percent. The PULSE has also been used to detect silicide phase transitions for Ti and Co reacted with silicon at different temperatures based on changes in sound velocity, density, and optical properties [91].

The commercially available MetaPULSe system uses a solid state, compact, ultrafast laser as a source for both the generating and detecting beams, dividing the single beam into two by means of a simple beam splitter. The sensitivity of the technique can be optimized to give improved throughput and sensitivity for specific materials, and a high throughput copper system is one example [91]. The MetaPULSe system is currently available in a highly automated stand alone configuration equipped with

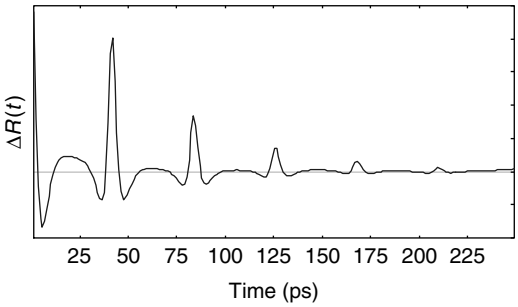


FIGURE 24.41 Simulated time dependent reflectivity of PULSE measurement of a TiN layer on a thick aluminum substrate. This data shows the effect of having a metal substrate instead of the inter-layer dielectric substrate shown in Figure 5.4. Data taken using a MetaPULSe picosecond laser ultrasonic sonar system. Figure courtesy Rob Stoner, Rudolph Technologies.

an optical microscope, pattern recognition software, and precision sample stage so that patterned wafers can be monitored. Because of its long working distance and high throughput, PULSE technology can also be applied to in situ measurements.

24.4.1.3 Resolution and Precision of Acoustic Film Thickness Measurements

The resolution of ISTS determined film thickness is nearly independent of film thickness and composition [90]. The resolution and precision of ISTS are better known for single layer films. Precision values of less than 0.2 nm are typically observed. The precision is sub-angstrom for a wide variety of single layer films. The reason for the small value for precision (greater measurement precision = smaller value of precision) is that the general shape of the ISTS data is a damped, high frequency oscillation. This general shape does not change with film thickness, only the frequency and damping characteristics change with thickness and composition. Since the oscillations can be averaged over many periods on every sample, precision values are always small.

The precision of PULSE measurements is related to the precision with which the echo times within a sample may be determined [91]. For a hypothetical measurement in which a single echo is analyzed, the precision is equal to the product of the sound velocity in the film (typically about 50 Å/ps) and the uncertainty in determining the centroid (time) of the echo $\Delta\tau$. The two most important factors which influence $\Delta\tau$ are the system noise (from mechanical, electrical, and optical fluctuations) and the linearity and precision of the mechanism used to establish the time base (i.e., the delay in time between the generating and measuring laser pulses). The latter is a mechanical delay line consisting of a retroreflector mounted on a linear stage, so that the time base linearity and precision are determined by the linear stage characteristics. In practice, a $\Delta\tau$ of about 0.006 ps can be achieved using readily available components over a range of echo times extending from 0 to over 2 ns. This gives a precision for thickness of order 0.3–0.6 Å for measurements based on a single echo, and this is typical of observed static measurement repeatability for films thicker than a few thousand Angstroms, giving a precision of order 0.1%, or better. For thinner films it is generally possible to observe several echoes within the finite delay range of the measuring system, and this gives a corresponding improvement in precision by a factor which is proportional to the number of echoes observed. Thus the precision in dimensions of length remains approximately constant for thicknesses ranging from a few hundred Angstroms to several microns. For very thin films the signal to noise ratio tends to degrade due to the increasing transparency of the films and the increasing overlap between successive echoes so that measurements become impractical for films thinner than about 25 Å (depending on the material). Therefore for films up to a few hundred Angstroms the measurement precision may vary from between 0.1 and 1% depending on the specific combination of optical properties and film thickness.

For multilayer films, the measurement precision is layer specific and also dependent on the proximity of each layer to the free surface as well as the ideality of the interfaces between layers. For multilayer samples with perfect interfaces between all layers, the measurement uncertainty (in dimensions of length) for any single layer increases relative to the equivalent single film case by an amount proportional to the number of films between it and the free surface. This increase in uncertainty is usually negligible since, in general, more than one echo is observed within a buried layer less than about 1000 Å thick.

24.4.1.4 Optical Measurement of Interconnect Dielectric and Metal Film Thickness

Ellipsometry can measure the thickness of very thin metal and metal-like films [98]. The depth of penetration of light is directly related to the imaginary part of the refractive index, k, of a material. The intensity of light at a depth, z, is given by: $I = I_0\, e^{-(4\pi kz/\lambda)}$. One way to estimate the maximum thickness of a metal film that ellipsometry can measure, presuming that the top layer is the only one being measured is to use the expression $d \sim \lambda/\pi k$ (5.16). Therefore, at $\lambda = 632.8$, the following thicknesses can be measured: TiN ~ 18 nm; Ti ~ 17 nm; amorphous Si ~ 107 nm; and Co 12 nm [98]. Since the requirements are for thicker films, ellipsometry is not in wide spread use for control of Ti and TiN deposition.

As mentioned in Section 24.3.2, SWE must be supplemented in order to measure film thickness for SiO_2 films thicker than ~283 nm. The values of Del and Psi repeat themselves after 283 nm making the measurement interpretation ambiguous. This problem has been overcome by use of multiple wavelengths, multi-angle, and SEs make the measurement unambiguous. The correct optical constants must be used for non-thermal SiO_2 films such as boron and phosphorous doped silica glass (BPSG) and boron and phosphorous doped TEOS.

Thickness measurements for low k dielectrics also require knowledge of the correct optical constants. Manufacturing process control for low κ dielectrics can be done even on some complicated film stacks [96]. Some of the proposed new dielectrics have different in-plane, η_{TE}, and out-of-plane, η_{TM}, refractive indices. These materials are characterized by a birefringence which is $\eta_{TE} - \eta_{TM}$. Some polyimide thin films show thickness dependence to their optical (and thermal, mechanical, and electrical) properties due to preferential orientation of the molecular chains in the plane of the film [99]. Ho and co-workers have published the dielectric constants (≤ 1 MHz) and the in-plane and out -of- plane refractive indices at $\lambda = 632.8$ nm [99]. As discussed in Section 24.3 of this chapter, the complex refractive index is the square root of the complex dielectric constant. At optical frequencies, the dielectric behavior is due to electric polarization while below 10 MHz the dielectric constant has contributions from electric and lattice polarization (Figure 24.41).

24.4.1.5 Metal Line Thickness and Copper Void Detection by Metal Illumination

The thickness of patterned metal lines can be measured in-line using Metal Illumination [95,100]. Borden has described this method which determines line resistance by measuring the thermal conductivity [95,100]. One can calibrate system response for un-patterned metal layers, as well as, measure the thickness of patterned metal lines. Resistance is a function of the line cross-section. The system used to measure metal line thickness as well as voids in patterned structures uses an 830 nm laser to inject heat into one or more metal lines [100]. The measurement is illustrated in Figure 24.42.

Visualizing the measurement process is useful for understanding its application to void determination. Borden describes the process in detail in Ref. [100]. Absorption of light from an 830 nm laser heats the metal lines within the spot diameter of 2 μm. The heating laser is modulated at kHz frequency. The heat flows away from this spot along the un-pattered film or along the lines away from the heated spot. This results in a temperature distribution. Under the laser spot used to heat the metal, the temperature is inversely proportional to the line cross-sectional area A and thermal conductivity K ($T \sim 1/(AK)$). Because the thermal and electrical conductivities of pure metals are related by the Wiedemann-Franz law, one can determine electrical conductivity and thus resistance if one measures the temperature of the metal lines [95]. The peak temperature varies as the resistance per unit length ($T \sim 1/(A\sigma) = \rho/A$), with σ the electrical conductivity and ρ the resistivity. A 980 nm laser is used to measure the reflectance of

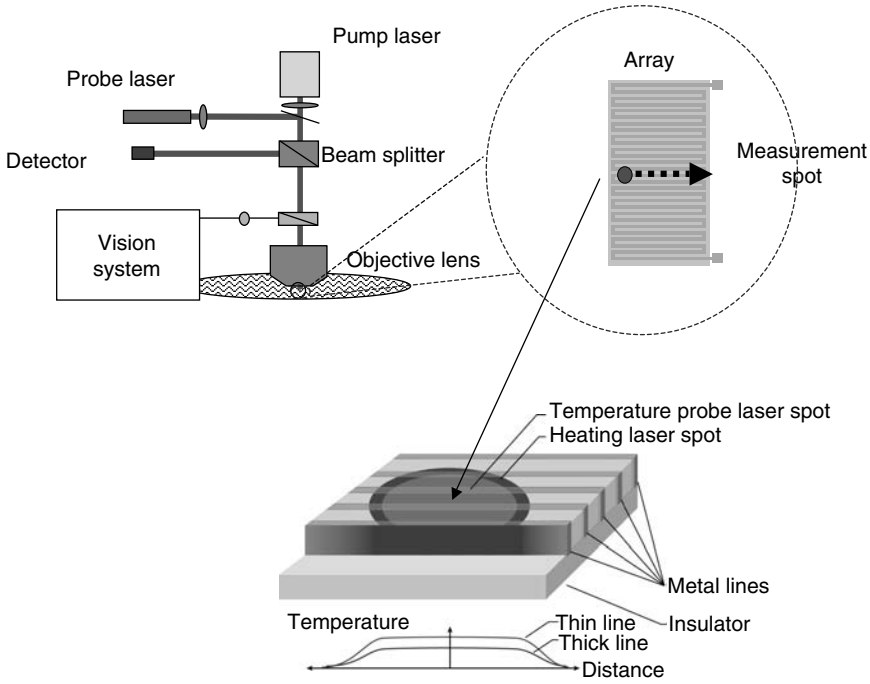


FIGURE 24.42 Metal illumination. Metal illumination can measure metal film and metal line thickness as well as detect voids in copper metal lines. The pump laser heats the metal film or metal lines, and the reflectivity of the probe laser light measures the resistivity of the metal. The resistivity value determines film or metal line thickness. Figure courtesy Peter Borden of Applied Materials.

copper as a function of temperature. The signal intensity will oscillate at the kHz frequency of the heating laser.

Detection of voids in copper metal lines and contact/via structures is a difficult task. Many of the proposed methods rely on the small change in volume of copper due to the presence of voiding. The challenge is to observe these volume changes when across wafer uniformity varies due to changes in metal line thickness. Variation in metal line thickness is a result of the different polishing rates of thick lines vs. dense lines vs. isolated lines. Metal illumination and ISTS have been applied to detection of voids in patterned copper metal structures.

In metal illumination, copper voids in metal lines are detected by the temperature change from the void induced reduction in the line cross-section. This effect has been experimentally verified using SEM cross-sections [100]. Voiding measurements are conducted by measuring a series of closely spaced sites (typically with site spacing smaller than the beam diameter) along a structure such as an isolated line or a via chain. This approximates a continuous scan along the line. In Reference [100], it is reported that metal illumination can detect a 1% voiding volume for voids that are 15% of the line cross-section down to 65 nm wide copper lines. At these dimensions, the void diameter is 52 nm with a density $N_{\text{Void}} = 1.9$ per microns for 90 nm wide lines and 37 nm diameter voids with a density of 2.7 per micron for 65 nm wide lines [100]. This density of voids per unit length is about equal one void in the measurement spot size of 2 μm. Ref. [100] also reports that the issues associated with CMP induced non-uniformity can be minimized by selecting the appropriate method of scanning.

Voids in contacts and vias can also be detected by the change in thermal conductivity that the void induces. When a good metal line or lines are scanned with metal illumination, the heat can be conducted away from the metal lines by the presence of a good via connection to the next metal level. Thus, the

temperature drops as one scans across the via. When the via has a void or is not connected to the metal line below, the temperature does not drop as much as it does for the good via [100]. In via measurements this method has an advantage over volume-based measurements. Structures such as via chains are designed with fat links and small vias in order to minimize the effect of a variation of link resistance. The MI method passes a current of heat through the via, and is thus predominantly sensitive to the via resistance change. Volume-based measurements will be equally sensitive to changes in link and via volume [100].

24.4.1.6 Pore Size Distribution in Porous Low *k*

Characterization of the distribution of pore sizes is a critical part of the analysis of porous films. Above, we described the use of x-ray reflectivity for measurement of film thickness. This method uses the specularly reflected x-rays (angle in = angle out), and it is an excellent method of measuring low *k* film thickness even for complicated stacks. As the x-rays transverse the low *k* film, the pores scatter x-rays away from the specular angle. This is also called diffuse scattering of x-rays by pores. Analysis of the intensity of the non-specularly scattered x-rays vs. the angle of incidence of the x-ray beam provides information on the pore size and the distribution of the pore size within the porous material [101]. The diffusely scattered x-rays need to be separated from the specularly reflected x-rays. There are a number of means of separating these signals. Moving the sample slightly away from the alignment used for x-ray reflectivity (XRR) allows use of existing XRR systems. Some suppliers have developed special detectors that capture diffuse scattering from a greater range of angles. The pore size distribution and average pore size is calculated from the scattering intensity vs. angle using well established x-ray theory and a model for the pore size distribution [101].

24.4.2 Ex-Situ Chemical Mechanical Polishing Process Control–Film Flatness and Quality

The purpose of CMP is to produce a flat surface over device topology for lithographic patterning. The CMP is used in both traditional metallization and Damascene (inlaid metal) processing. In traditional metallization processes, metal is deposited, patterned, etched, and overcoated with oxide which is polished flat. In Damascene (inlaid metal) processes, oxide (or low *k* dielectric) is deposited, patterned, and etched, and metal fill is deposited and then polished back to oxide. The CMP must be monitored for

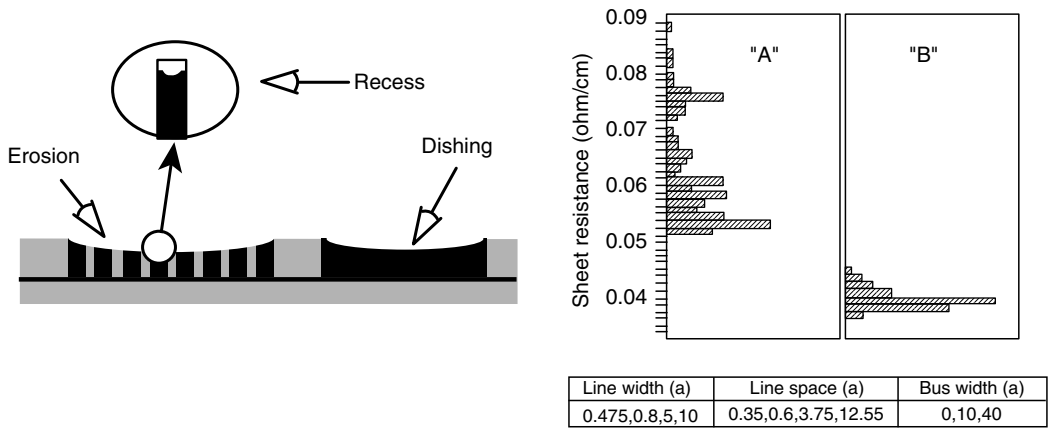


FIGURE 24.43 Defects in inlaid metal CMP processes. Figure from Farkas, J., and M. Freedman, In *Proceedings of the Fourth Workshop on Industrial Applications of Scanned Probe Microscopy*, NIST, May, 1997. With permission.

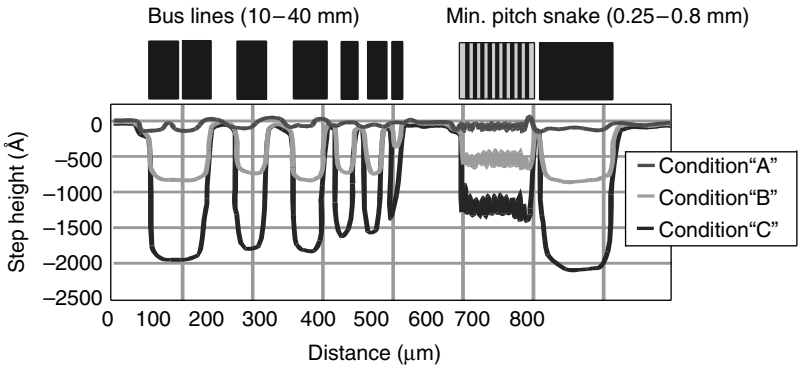


FIGURE 24.44 Comparison of CMP process conditions using high resolution profilometry. Figure from Farkas, J., and M. Freedman, In *Proceedings of the Fourth Workshop on Industrial Applications of Scanned Probe Microscopy*, NIST, May, 1997. With permission.

both local flatness within a die and across the wafer. The CMP can generate some radial variation in removal across the wafer. Both traditional and Damascene process are adversely effected by this radial variation in removal rate. Over etching of contact and via structures at the edge of the wafer can result when an oxide layer over metal is over thinned toward the wafer edge. In this case, optical film thickness is used to monitor oxide CMP for traditional metallization processes to ensure that differences in thickness of the oxide layer do not result in over etching during fabrication of contact/via openings. Stylus profilometry or scanned probe microscopy are used to determine film flatness after CMP.

The local topography across a die is known to induce non-uniform polishing, and particles in the slurry used during CMP can either be left on the surface or produce scratches that can result in yield loss. Scratches, pits, and residual particles have been referred to as localized irregularities [102]. Dishing, erosion, and recess result in insufficient planarity [103]. In Figure 24.43 we show some of the types of polishing defects that result in insufficient planarity. Process development requires careful monitoring of test structures. In Figure 24.44, we show an example of the recess of a CMP processed test structure under 3 different polishing conditions [103]. The test structure has several different types of metal lines: bus lines, bond pads, and minimum feature size metal lines. The CMP of layers containing contact and via

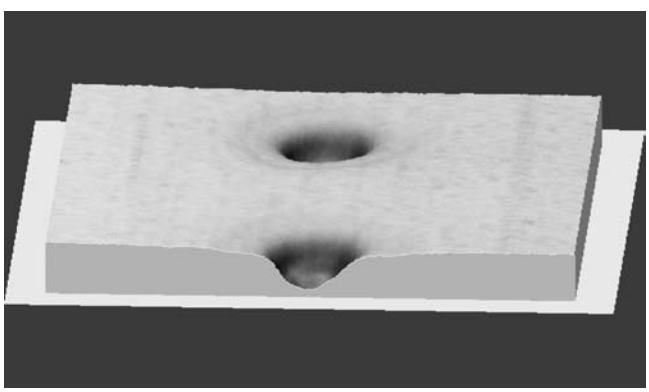


FIGURE 24.45 Profilometry characterization of CMP of via and contact structures. Figure courtesy Jason Schneir (KLA-Tencor).

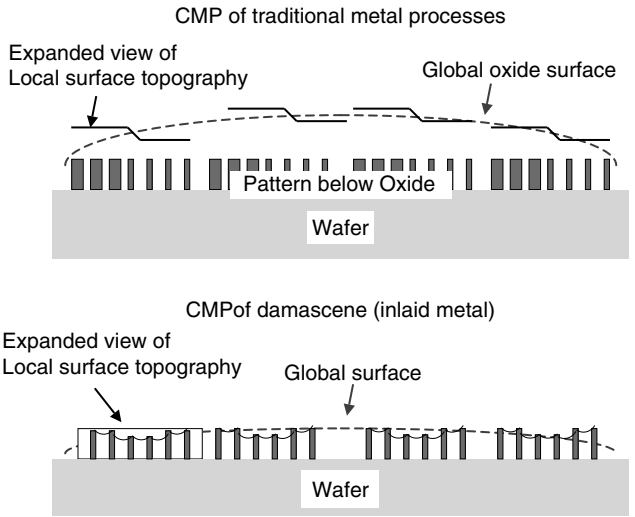


FIGURE 24.46 Process monitoring requirements for CMP of tradition metal and damascene (inlaid metal) processes. The local topography must be monitored over the area of lithographic exposure, while the radial variation of thickness must also be monitored. Optical thickness measurements can be used to determine the local variation of oxide thickness in traditional CMP to insure that via/contact etch processes do not "over etch".

structures must also be monitored during process development. Local scans are used to characterize the recess of tungsten plugs after CMP as shown in Figure 24.45. Issues associated with local flatness and global over-polishing is shown in Figure 24.46 for both traditional and Damascene processes.

Stylus profilometers and the new, high resolution profilometers provide information on planarity both locally and globally. The lithographic exposure tool images the reticle mask over the distance of a die ($\sim$25 mm$\times$25 mm) with a limited depth of focus ($<$600 nm using 248 nm exposure tools for 180 nm technology generations). The KLA-Tencor HRP-200 uses a diamond probe tip with a tip radius of 0.05 mm and a scan length of up to 205 mm. A lateral resolution of 1 nm is possible when local areas are analyzed [102]. The step height repeatability is 0.8 nm for features that are 1 μm in height.

Maps from wafer flatness tools (such as profilometers) may become part of the lot by lot database of metrology data management systems. In-line process control for CMP is done using test structures which are processed after a change in recipe or polishing pads on the CMP tool.

24.5 In-FAB FIB

Focused ion beam systems are now available in two versions, with dual electron beam and ion beam columns and as ion beam systems only [104]. FIB can be supplemented with a SIMS and the dual column systems can also be equipped with EDS. FIB has been used in-FAB for control of trench capacitor and contact/via structures. FIB cross-sections of these structures allow detailed imaging of critical titanium nitride

Acknowledgments

I thank and acknowledge Jimmy Price, P.Y. Hung, Ben Bunday, and Hugo Celio for their assistance with revising this chapter. I gratefully acknowledge the review and figures from Peter Borden (Carrier and Metal Illumination) and discussion and review by Michael Gostein (ISTS). I also acknowledge those that helped with the chapter that appeared in the first addition of this volume: Arnie Ford, Dan Holladay,

Hershel Marchman, Kevin Monahan, Chris Nelson, Paul Tobias, Harland Tompkins, George Collins, George Brown, Steve Weinzeirl, John Hauser, Jimmy Wortman, Mathew Banet, John Hanselman, Gary Schwartz, and Randy Goodall.

References

1. International Technology Roadmap for Semiconductors. Semiconductor Industry Association, 2003.
2. Eastman, S. A. "Evaluating Automated Wafer Measurement Instruments." *SEMATECH Technology Transfer Document 94112638A-XFR*. 1995. A PDF file can be downloaded from SEMATECH's public web site at http://www.sematech.org/public/docubase/abstracts/wrapper26.htm
3. Ballard, D. H., D. W. McCormack, Jr., T. L. Moore, M. Pore, J. Prins, and P. A. Tobias. "A Comparison of Gauge Study Practices." *Proceedings From the 1997 Joint Statistical Meetings of the American Statistical Association, Quality and Productivity Section*. 1998.
4. "SEMI E89 Guide for Measurement System Analysis." 3081. Zanker Road, San Jose, CA 95134: SEMI Standards.
5. Currie, L. A. "Limits for Quantitative Detection and Quantitative Determination." *Anal. Chem.* 40 (1968): 586–93.
6. Zeitzoff, P. M. "Modeling of Statistical Manufacturing Sensitivity and of Process Control and Metrology Requirements for a 0.18-μm NMOSFET." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 117–41. New York: Dekker, 2001.
7. Zeitzoff, P. M., A. F. Tasch, W. E. Moore, S. A. Khan, and D. Angelo. "Modeling of Manufacturing Sensitivity and of Statistically Based Process Control Requirements for a 0.18 μm NMOS Device." In *Characterization and Metrology for ULSI Technology*, edited by D. G. Seiler, A. C. Diebold, W. M. Bullis, T. J. Shaffner, R. McDonald, and E. J. Walters, 73–82. New York: AIP Press, 1998.
8. Gaitonde, D., and D. M. H. Walker. "Test Quality and Yield Analysis Using DEFAM Defect to Fault Mapper." *Proceedings of the IEEE International Conference on CAD 1993*, 78–83, 1993.
9. Khare, J., and W. Maly. "Rapid Failure Analysis Using Contamination-Defect-Fault (CDF) Simulation." *Proceedings of the Fourth IEEE/UCS/SEMI International Symposium on Semiconductor Manufacturing 1995 (ISSM '95)*, IEEE Catalogue Number 95CH35841, 136, 1995.
10. Diebold, A. C., ed. *Handbook of Silicon Semiconductor Metrology*. New York: Marcel Dekker, 2001.
11. Joy, D. Private communication.
12. Marchman, H. M. "Scanning Electron Microscope Matching and Calibration for Critical Dimensions." *Future FAB Int.* 3 (1997): 345–54.
13. Goldstein, J. I., D. E. Newbury, P. Echlin, D. C. Joy, C. Lyman, P. E. Echlin, E. Lifshin, L. Sawyer, and J. Michael. "Electron Beam–Specimen Interaction." In *Scanning Electron Microscopy and X-Ray Microanalysis* 3rd ed., 689. New York: Kluwer Academic/Plenum, 2003.
14. Postek, M. T., and A. E. Vladar. "Critical Dimension Metrology and the Scanning Electron Microscope." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 295–334. New York: Marcel Dekker, 2001.
15. Villarrubia, S., A. E. Vladar, and M. T. Postek. "A Simulation Study of Repeatability and Bias in the CD-SEM." In *Proceedings of SPIE* 5038, 138–49, 2003.
16. Mayer, J., K. Huizenga, E. Solecky, C. Archie, G. W. Banks, R. Cogley, C. Nathan, and J. Robert. "New Apparent Beam Width Artifact and Measurement Methodology for CD-SEM Resolution Monitering." *APIE* 5038, 699–710, 2003.
17. Bunday, B. D., M. Bishop, J. R. Swyers, and K. Lensing. "Quantitative Profile-Shape Measurement Study on a CD-SEM with Application to Etch-Bias Control and Several Different CMOS Features." In *Proceedings SPIE* 5038, 383–95, 2003.
18. Sullivan, N., R. Dixson, B. Bunday, and M. Mastovich. "Electron Beam Metrology of 193 nm Resists at Ultra-low Voltage." In *Proceedings of SPIE* 5038, 483–92, 2003; Sullivan, N., M. Mastovich, S. Bowdoin, and R. Brandon. "CD-SEM Acquisition Effects on 193 Resist Line Slimming." *Proceedings of SPIE* 5038, 618–23, 2003.

19. Bunday, B. Private communication.
20. Bunday, B. et al. *Unified Advanced CD-SEM Specification for sub 130 nm Technology*. This document is available at www.sematech.org
21. Bunday, B., and M. Davidson. "Use of Fast Fourier Transform Methods in Maintaining Stability of Production CD-SEMs". *In Proceedings of SPIE* 3998, 913–22, 2000.
22. Cresswell, M. W., J. J. Sniegowski, R. N. Ghothtagore, R. A. Allen, W. F. Guthrie, A. W. Gurnell, L. W. Lindholm, R. G. Dixson, and E. C. Teague. "Recent Developments in Electrical Linewidth and Overlay Metrology for Integrated Circuit Fabrication Processes." *Jpn. J. Phys.* 35 (1996): 6597–609.
23. Martin, Y., and H. K. Wickramasinghe. "Precision Micrometrology." *Future FAB Int.* 1 (1996): 253–60.
24. Martin, Y., and H. K. Wickramasinghe. "Method for Imaging Sidewalls by Atomic Force Microscopy." *Appl. Phys. Lett.* 64 (1994): 2498–500.
25. Feenstra, R. M., and J. E. Griffith. "Semiconductor Characterization with Scanning Probe Microscopies." In *Semiconductor Characterization: Present Status and Future Needs*, edited by W. M. Bullis, D. G. Seiler, and A. C. Diebold, 295–307. New York: AIP, 1996.
26. Griffith, J. E., and D. A. Grigg. "Dimensional Metrology with Scanning Probe Microscopes." *J. Appl. Phys.* 74 (1993): R83–109.
27. Griffith, J. E., H. M. Marchman, and L. C. Hopkins. "Edge Position Measurement with a Scanned Probe Microscope." *J. Vac. Sci. Technol.* B12 (1994): 3567–70.
28. Marchman, H. M., J. E. Griffith, J. Z. Y. Guo, and C. K. Cellar. "Nanometer-Scale Dimensional Metrology for Advanced Lithography." *J. Vac. Sci. Technol.* B12 (1994): 3585–90.
29. Marchman, H. M., and J. E. Griffith. "Scanned Probe Microscope Dimensional Metrology." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 335–76. New York: Marcel Dekker, 2001.
30. Standard Practice for Measuring and Reporting Probe Tip Shape in Scanning Probe Microscopy, E1813-96, American Society for Testing and Materials.
31. Dixson, R., A. Guerry, M. Bennett, T. Vorburger, and M. Postek. "Toward Traceability for At-Line AFM Dimensional Metrology." *Metrology, Inspection, and Process Control for Microlithography, Proceedings of SPIE* 4689, 313–35, 2002.
32. Raymond, C. J. "Scatterometry for Semiconductor Metrology." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 477–513. New York: Marcel Dekker, 2001.
33. Opsal, J., H. Chu, Y. Wen, Y. C. Chang, and G. Li. "Fundamental Solutions for Real-Time Optical CD Metrology." *Metrology, Inspection, and Process Control for Microlithography, Proceedings of SPIE* 4689, 163–76, 2002.
34. Conrad, E. W., and D. P. Paul. Method and Apparatus for Measuring the Profile of Small Repeating Lines. U.S. Patent 5, 963, 329; Moharam, M. G., T. K. Gaylord, G. T. Sinerbox, H. Werlich, and B. Yung. "Diffraction Characteristics of Photoresist Surface-Relief Grating." *Appl. Opt.* 23 (1984): 3214–20.
35. Sendelbach, M., and C. Archie. "Scatterometry Measurement Precision and Accuracy below 70 nm." *Metrology, Inspection, and Process Control for Microlithography, Proceedings of SPIE* 5038, 224–38, 2003.
36. Sonderman, T., M. Miller, and C. Bode. "APC as a Competitive Manufacturing Technology: *Getting It Right for 300 mm*." In *AEC/APC Symposium XIII, International SEMATECH*, 2001.
37. Petersen, J. S., J. D. Byers, and R. A. Carpio. "The Formation of Acid Diffusion Wells in Acid Catalyzed Photoresists." *Microelec. Eng.* 35 (1997): 169–74.
38. Lindholm, L. L., R. A. Allen, and M. W. Cresswell. "Microelectronic Test Structures for Feature Placement and Electrical Linewidth Metrology." In *Handbook of Critical Dimension Metrology and Process Control*, edited by K. M. Monahan, 91–132. Bellingham: SPIE Optical Engineering Press, 1993.
39. Chain, E. E., M. D. Griswold, and B. P. Singh. "In-Line Electrical Probe for CD Metrology." *Process Equipment and Materials Control in Integrated Circuit Manufacture II. SPIE Proceedings 2876*, edited by A. Iturraldo and T. Lin. 135–46, 1996.

40. Lee, K., M. Shur, T. A. Fjeldly, and T. Ytterdal. "Basic MOSFET Theory". In *Semiconductor Device Modeling for VLSI*, 240. Englewood Cliffs: Prentice-Hall, 1993.
41. Sullivan, N. T. "Semiconductor Pattern Overlay." In *Handbook of Critical Dimension Metrology and Process Control*, edited by K. M. Monahan, 160–88. Bellingham: SPIE Optical Engineering Press, 1993.
42. Starikov, A. "Metrology of Image Placement." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 411–76. New York: Marcel Dekker, 2001.
43. Tompkins, H. G. *A User's Guide to Ellipsometry*. New York: Academic Press, 1993.
44. Jellison, G. E. "Physics of Optical Metrology of Silicon-Based Semiconductor Devices." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 723–60. New York: Dekker, 2001.
45. Opsal, J., J. Fanton, J. Chen, J. Leng, L. Wei, C. Uhrich, M. Senko, C. Zaiser, and D. E. Aspnes. *Broadband Spectral Operation of a Rotating-Compensator Ellipsometer.*
46. System description.
47. System description.
48. Diebold, A. D., D. Venables, Y. Chabal, D. Muller, M. Welden, and E. Garfunkel. "Characterization and Production Metrology of Thin Transistor Gate Oxide Films." *Mater. Sci. Semicon. Process.* 2 (1999): 103–47.
49. Hayzelden, C. "Gate Dielectric Metrology." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 17–48. New York: Dekker, 2001.
50. Chandler-Horowitz, D., and G. A. Candela. "On the Accuracy of Ellipsometric Thickness Determination for Very Thin Films." *J. De Physique* C10 (1983): 23–6.
51. Fang, S. J., and C. R. Helms. "Ellipsometric Model Studies of the Si Surface." In *Contamination Control and Defect Reduction in Semiconductor Manufacturing III*. 94-9, 267–76. PV: Electrochemical Society, 1994.
52. Nguyen, N. V., D. Chandler-Horowitz, P. M. Amirtharaj, and J. G. Pellegrino. "Spectroscopic Ellipsometry Determination of the Properties of the Thin Underlying Strained Si Layer and the Roughness at Si/SiO_2 Interfaces." *Appl. Phys. Lett.* 64 (1994): 5599.
53. Fang, S. J., W. Chen, T. Yamanaka, and C. R. Helms. "Influence of Interface Roughness on Silicon Oxide Thickness Measured by Ellipsometry." *J. Electrochem. Soc.* 144 (1997): L231–3.
54. Nicollian, E. H., and J. R. Brews. Metal Oxide Silicon Capacitor at Low Frequencies. In *MOS (Metal Oxide Semiconductor) Physics and Technology*, 71–98. New York: Wiley, 1982.
55. Nicollian, E. H., and J. R. Brews. "Metal Oxide Silicon Capacitor at Intermediate and High Frequencies." In *MOS (Metal Oxide Semiconductor) Physics and Technology*, 99–175. New York: Wiley, 1982.
56. Blood, P., and J. W. Orton. "Capacitance–Voltage Profiling." *The Electrical Characterization of Semiconductors: Majority Carriers and Electron States*, 220–65. New York: Academic Press, 1992.
57. Emerson, N. G., and B. J. Sealy. "Capacitance–Voltage and Hall Effect Measurements." In *Analysis of Microelectronic Materials and Devices*, edited by M. Grasserbauer, and H. W. Werner, 865–85. New York: Wiley, 1991.
58. Schroder, D. K. "Oxide and Interface Trapped Charge." *Semiconductor Material and Device Characterization*, 244–96. New York: Wiley, 1990.
59. Lee, K., M. Shur, T. A. Fjeldly, and T. Ytterdal. "Surface Charge and the Metal Insulator Semiconductor Capacitor." *Semiconductor Device Modeling for VLSI*, 196–228. Englewood Cliffs: Prentice-Hall, 1993.
60. Vogel, Eric M., and VeenaMisra. "MOS Device Characterization." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 59–96. New York: Dekker, 2001.
61. Larson, L. "Metrology for Ion Implantation." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 49–58. New York: Dekker, 2001.
62. Hauser, J. R. and K. Ahmed. Characterization of Ultra-Thin Oxides Using Electrical C–V and I–V Measurements. In *Characterization and Metrology for ULSI Technology*, 449, 235–39. New York: AIP Conference Proceedings, 1998.

63. Vogel, E. M. "Issues with Electrical Characterization of Advanced Gate Dielectrics." In *Metal-Oxide-Semiconductor Devices, Proceedings* WoDIM 2002, edited by IMEP, Grenoble, France, 2002.
64. Weinzeirl, S. Private Communication.
65. Briggs, D. "XPS: Basic Principles, Spectral Features, and Quantitative Analysis." In *Surface Analysis by Auger and X-Ray Photoelectron Spectroscopy*, edited by D. Briggs, and J. T. Grant, 31–56. UK: IM Publications and Surface Spectra Limited, 2003.
66. Cumpson, P. J. "Angle Resolved X-Ray Photoelectron Spectroscopy." In *Surface Analysis by Auger and X-Ray Photoelectron Spectroscopy*, edited by D. Briggs, and J. T. Grant, 651–76. UK: IM Publications and Surface Spectra Limited, 2003.
67. Drummond, I. W. "AES Instrumentation and Performance." In *Surface Analysis by Auger and X-Ray Photoelectron Spectroscopy*, edited by D. Briggs, and J. T. Grant, 117–44. UK: IM Publications and Surface Spectra Limited, 2003.
68. Goldstein, J. I., D. E. Newbury, P. Echlin, D. C. Joy, A. D. Romig, Jr., C. E. Lyman, C. Fiori, and E. Lifshin. "X-Ray Spectral Measurement: WDS and EDS." In *Scanning Electron Microscopy and X-Ray Microanalysis*, 2nd ed., 273–340. New York: Plenum, 1992.
69. Larson, L. "Metrology for Ion Implantation." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold. New York: Dekker, 2001.
70. Yarling, C. B., and M. I. Current. "Ion Implantation Process Measurement, Characterization, and Control." In *Ion Implantation Science and Technology*, edited by J. F. Zeigler, 674–721. Poughkeepsie: Ion Implantation Technology Co., 1996.
71. Larson, L., and M. I. Current. "Doping Process Technology and Metrology." In *Characterization and Metrology for ULSI Technology*, edited by D. G. Seiler, A. C. Diebold, M. Bullis, R. McDonald, and T. J. Shaffner, 143–52. New York: AIP, 1998.
72. Schroder, D. K. "Resistivity". In *Semiconductor Material and Device Characterization*, 1–40. New York: Wiley, 1990.
73. Johnson, W. H. "Sheet Resistance Measurements of Interconnect Films." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 215–44. New York: Dekker, 2001.
74. Smith, W. L., A. Rosenwaig, and D. L. Willenbourg. *Appl. Phys. Lett.* 47 (1985): 584.
75. Schroeder, D. K., B. Schueler, and G. S. Strossman. "Electrical, Physical, and Chemical Characterization." Chap. 28, (this volume).
76. Dowsett, M. G. "SIMS Depth Profiling of Ultra Shallow Implants and Junctions in Silicon—Present Performance and Future Potential." In *Secondary Ion Mass Spectrometry SIMS XI*, 259–64. New York: Wiley, 1998.
77. Dowsett, M. G., T. J. Ormsby, D. T. Elliner, and G. A. Cooke. "Establishment of Equilibrium in the Top Nanometers Using Sub-KeV Beams." In *Secondary Ion Mass Spectrometry SIMS XI*. 371–8. New York: Wiley, 1998.
78. Dowsett, M. G., G. A. Cooke, D. T. Elliner, T. J. Ormsby, and A. Murrell. "Experimental Techniques for Ultra-Shallow Profiling Using Sub-keV Primary Ion Beams." In *Secondary Ion Mass Spectrometry SIMS XI*, 285–8. New York: Wiley, 1998.
79. Borden, P., L. Bechtler, K. Lingel, and R. Nijmeijer. "Carrier Illumination of Ultra-Shallow Implants." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 97–116. New York: Marcel Dekker, 2001.
80. Borden, P., P. Gillespie, A. Al-Bayati, and C. Lazik. "In-Line Implant/Anneal Module Monitoring of Ultra-Shallow Junctions." *Proceedings of the 14th International Conference on Ion Implant Technology*, IEEE, CD Rom. 9no paper version yet.
81. Zeitzoff, P. M., J. A. Hutchby, and H. R. Huff. "MOSFET and Front-End Process Integration: Scaling Trends, Challenges, and Potential Solutions through the End of the Roadmap." *Int. J. High-Speed Electron. Syst.* 12 (2002): 267–93.
82. Åberg, O., O. Olubuyide, C. Ní Chléirigh, I. Lauer, D. A. Antoniadis, J. Li, R. Hull, and J. L. Hoyt. "Electron and Hole Mobility Enhancements in Sub-10 nm-Thick Strained Silicon Directly on Insulator Fabricated by a Bond and Etch-Back Technique." *Digest of 2004 Symposium on VLSI Technology*, 52–3, (IEEE Catalog no. 04CH37571).

83. Thompson, S. E., G. Sun, K. Wu, J. Lim, and T. Nishida. "Key Differences for Process-Induced Uniaxial vs. Substrate-Induced Biaxial Stressed Si and Ge Channel MOSFETs." *IEDM Tech. Digest* (2004): 221–4.
84. Ghani, T., M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, et al. *IEDM Tech. Digest* (2003): 978–80, (IEEE Catalog no. 03CH37457).
85. Giles, D., M. Armstrong, C. Auth, S. M. Cea, T. Ghani, T. Hoffmann, R. Kotlyar, et al. "Understanding Stress Enhanced Performance in Intel 90 nm CMOS Technology." *2004 Symposium on VLSI Technology Technical Digest of Papers*, 118–9.
86. Chidambaram, P. R., B. A. Smith, L. H. Hall, H. Bu, S. Chakravarthi, Y. Kim, A. V. Samoilov, et al. "35% Drive Current Improvement from Recessed-SiGe Drain Extensions on 37 nm Gate Length PMOS." *Digest of 2004 Symposium on VLSI Technology*, 48–49.
87. Pidin, S., T. Mori, K. Inoue, S. Fukuta, N. Itoh, E. Mutoh, K. Ohkoshi, et al. "A Novel Strain Enhanced CMOS Architecture Using Selectively Deposited High Tensile and High Compressive Silicon Nitride Films." *IEDM Techn. Digest* (2004): 213–6, (IEEE Catalog no. 04CH37602).
88. Bianchi, R. A., G. Bouche, and O. Roux-dit-Buisson. "Accurate Modeling of Trench Isolation Induced Mechanical Stress Effects on MOSFET Electrical Performance." *IEDM Techn. Digest* (2002): 117–20, (IEEE Catalog no. 02CH37358).
89. Xiang, Q., J-S. Goo, J. Pan, B. Yu, S. Ahmed, J. Zhang, and M.-R Lin. "Strained Silicon NMOS with Nickel–Silicide Metal Gate." *2003 Symposium on VLSI Technology Digest of Technical Papers*, 101–2.
90. Gostein, M., M. Banet, M. A. Joffe, A. A. Maznev, R. Sacco, J. A. Rogers, and K. A. Nelson. "Thin Film Metrology Using Impulsive Stimulated Thermal Scattering." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 167–96. New York: Marcel Dekker, 2001.
91. Diebold, A. C., and R. Stoner. "Metal Interconnect Process Control Using Picosecond Ultrasonics." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 197–214. New York: Marcel Dekker, 2001.
92. Johnson, W. H. "Sheet Resistance Measurements of Interconnect Films." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 215–44. New York: Marcel Dekker, 2001.
93. Deslattes, R. D., and R. J. Matyi. "Analysis of Thin Layer Structures by X-Ray Reflectometry." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 789–810. New York: Marcel Dekker, 2001.
94. Lachance, G. R., and F. Claisse. "Thin Films." In *Quantitative X-Ray Fluorescence Analysis: Theory and Practice*, 211–16. New York: Wiley, 1995; Jenkins, R., R. W. Gould, and D. Gedcke. *Quantitative X-Ray Spectrometry*. New York: Marcel Decker, 1981.
95. Borden, P., J. Madsen, and J. P. Li. "Non-Destructive, High-Resolution Metrology of Fine Metal Arrays." *Future FAB*10 (2000): 261–5; Wu, C. M., M. Y. Wang, C. T. Lin, C. W. Chang, M. H. Tsai, C. H. Hsieh, S. L. Shue, et al. "Non-Destructive In-Line Cu/Low-K Measurement Using Metal Illumination Method," to be published at the 2003 International Interconnect Technology Conference, Burlingame, CA, June 2–4, 2003.
96. Diebold, A. C., W. W. Chism, T. G. Dziura, and A. Kanan. "Metrology for On-Chip Interconnect Dielectrics." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 149–66. New York: Marcel Dekker, 2001.
97. Matriz 100 Reference.
98. Asinovshy, L. Limits of Thickness Measurement. In Rudolph Technologies Applications Note, August 9, 1994.
99. Kiene, M., M. Morgan, J. H. Zhao, C. Hu, T. Co, and P. S. Ho. "Characterization of Low Dielectric Constant Materials." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 245–78. New York: Marcel Dekker, 2001.
100. Borden, P. A., J. P. Li, S. Smith, A. C. Diebold, and W. Chism. "Line and Via Voiding Measurements in Damascene Copper Lines Using Metal Illumination." *IEEE Trans. Semicond. Manufact.* 16, (2003): 409–16.
101. Rauscher, M., et al. *Phys. Rev. B* 52 (1995): 16855.

102. Mathai, A., and C. Hayzelden. "High-Resolution Profilometry for CMP and Etch Metrology." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 279–94. New York: Marcel Dekker, 2001.
103. Farkas, J., and M. Freeman. "New Requirements for Planarity and Defect Metrology in Soft Metal CMP." In *Proceedings of the Fourth Workshop on Industrial Applications of Scanned Probe Microscopy.* NIST, May, 1997.
104. Diebold, A. C., and R. McDonald. "The At-Line Characterization Laboratory of the 90s: Characterization Laboratories (FAB-LABS) Used to Ramp-Up New FABS and Maintain High Yield." *Future FAB Int.* 1, no. 3 (1997): 323–30.

25

In-Situ Metrology

25.1 Introduction **25**-1

25.2 Process State Sensors **25**-4
Temperature • Gas Phase Reactant Concentration • RF Properties • Wall Deposition Sensor

25.3 Wafer-State Sensors **25**-31
Film Thickness and Uniformity • Feature Profile

25.4 Measurement Techniques for Potential Sensors..... **25**-47
Ellipsometry • Epi Resistivity and Thickness

25.5 Software for In-Situ Metrology **25**-53
Data Collection Software • FDC Analysis Software • Model-Based Process Control Software

25.6 Use of In-Situ Metrology in SC Manufacturing **25**-55
Fault Detection and Classification • Fault Interdiction • Fault Prognosis • Model-Based Process Control

References **25**-57

Gabriel G. Barna
Texas Instruments, Inc.

Brad VanEck
SEMATECH

25.1 Introduction

Since the early 1960s, semiconductor manufacturing has historically relied on statistical process control (SPC) for maintaining processes within prescribed specification limits. This is fundamentally a passive activity based on the principle that the process parameters—the hardware settings—be held invariant over long periods of time. The SPC then tracks certain unique, individual metrics of this process—typically some wafer-state parameter—and declares the process to be out-of-control when the established control limits are exceeded with a specified statistical significance. While this approach has established benefits, it suffers from (a) its myopic view of the processing domain—looking at one, or only a few parameters and (b) its delayed recognition of a problem situation—looking at metrics generated only once in a while or with a significant time delay relative to the rate of processing of wafers.

In the late 1990s, while semiconductor manufacturing continues to pursue the ever-tightening specifications due to the well-known problems associated with the decreasing feature size and increased wafer size, it became clear that both these constraints have to be removed in order to stay competitive in the field. Specific requirements are that

- processing anomalies be determined by examining a much wider domain of parameters;
- processing anomalies be detected in shorter timeframes; within-wafer or at least wafer-to-wafer;
- processing emphasis be focused on decreasing the variance of the wafer-state parameters instead of controlling the variance of the set points.

Advanced process control (APC) is the current paradigm that attempts to solve these three specific problems. Under this methodology, the fault detection and classification (FDC) component addresses

the first two requirements, model-based process control (MBPC) addresses the last one. In contrast to the SPC methodology, APC is a closed-loop, interactive method where the processing of every wafer is closely monitored in a time scale that is much more relevant (within-wafer or wafer-to-wafer) to the manufacturing process. When a problem is detected, the controller can determine whether to adjust the process parameters (for small deviations from the normal operating conditions) or to stop the misprocessing of subsequent wafers (for major deviations from the standard operating conditions).

The APC paradigm is a major shift in operational methods and requires a complex, flexible architecture to be in place to execute the above requirements. A schematic representation of this architecture is provided in Figure 25.1. Briefly, this system starts with the processing tool and sets of in-situ sensors and ex-situ metrology tools to provide data on the performance of the tool. When the performance exceeds some pre-defined specifications, actions can be taken to either terminate the processing or reoptimize the settings of the equipment parameters via the model tuner and the pertinent process models. The purpose of this brief APC overview was to provide the context for this chapter on in-situ metrology. In-situ sensors are becoming more widespread, as they are one of the key components of this APC paradigm.

The goal of this chapter is to detail the fundamentals of in-situ process and wafer-state sensors in original equipment manufacturer (OEM) tools, and their use in APC, as this is the path that semiconductor manufacturing now has to aggressively pursue. This message is clearly articulated in the 1997 version of the National technology roadmap for semiconductors (NTRS), which states: "To enable this mode of operation (APC), key sensors are required for critical equipment, process, and wafer-state parameters. It is essential that these sensors have excellent stability, reliability, reproducibility, and ease of use to provide high quality data with the statistical significance needed to support integrated manufacturing" [1].

Hence, this chapter will provide information for the in-situ sensors that are *commercially available* (i.e., not test-bed prototypes) and are currently, or soon will be, used in OEM tools for the measurement and control of process state and wafer-state properties. In-situ sensors are those that monitor the process state of the tool or the state of the wafer during the processing of each wafer. For the sake of completeness, in-line sensors will also be included, as some of the sensor technologies are only applicable in this format. In-line sensors measure wafer state in some location close to the processing, such as a cool-down station in a deposition reactor or a metrology module on a lithography track system. Metrology tools that are currently used off-line, but are seen to be moving towards a simpler in-situ or in-line sensor embodiment

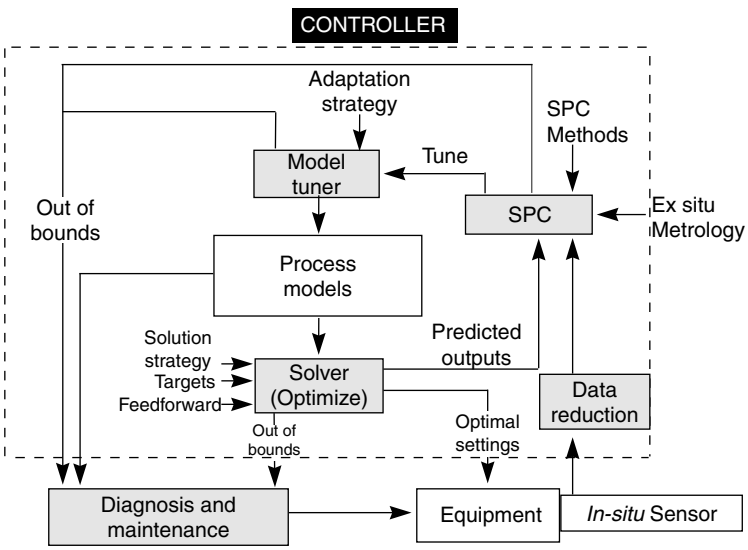


FIGURE 25.1 Architecture for advanced process control.

(e.g., ellipsometer, Fourier transform infrared (FTIR) film thickness) will be included. Sensors used for tool development (e.g., intrusive RF probes, spatially resolved spectroscopy) are not included. Sensors used for gas flow and pressure control are also omitted, as these are well-established technologies; while in-situ particle sensors are covered in the chapter on contamination free manufacturing. For each sensor described, the following information will be included based on input from the sensor manufacturer:

- the fundamental operating principle behind each sensor, with greater detail for the less-common ones;
- practical issues in the use and interfacing of these sensors.

When a number of manufacturers exist for a given sensor, references will be provided to several manufacturers; although there is no claim that this list will be totally inclusive. The sensors included in this chapter are ones that provide most of the features of an ideal in-situ sensor. These features are low

TABLE 25.1 Web Links to Major Companies Manufacturing Sensors for Semiconductor Manufacturing

Sensor Company	Web Site	Product
Nanometrics	http://www.nanometrics.com	Scatterometer, Fourier transform infrared (FTIR)
Advanced Energy	http://www.advanced-energy.com	RF systems, sensors
Digital Instruments	http://www.di.com	AFM for chemical–mechanical polishing (CMP) (offline)
ENI	http://www.enipower.com	RF probes
Ferran Scientific	http://www.ferran.com/main.html	Residual gas analysis (RGA)
Ircon	http://www.designinfo.com/ircon/html	Noncontact IR thermometers
KLA-Tencor	http://www.kla-tencor.com	Stress, thin film measurement, wafer inspection, metrology
Leybold Inficon, Inc.	http://www.inficon.com	Residual gas analyzer (RGA), leak detectors, full wafer interferometry, quartz crystal microbalance (QCM)
Lucas Labs	http://www.LucasLabs.com	Plasma diagnostics
Luxtron	http://www.luxtron.com	Endpoint for plasma and CMP
Ocean Optics	http://www.oceanoptics.com/homepage.asp	Spectrometer-on-a-card
Nova Measuring Instruments	http://www.nova.co.il	Spectrophotometric integrated thickness monitors for CMP
On-Line Technologies, Inc.	http://www.online-ftir.com	FTIR spectrometer for wafer state and gas analysis
Panametrics	http://www.industry.net/panametrics	Moisture, O_2 analyzers
Princeton Instruments	http://www.prinst.com	Imaging, spectroscopy
Quantum Logic Corporation	http://www.quantumlogic.com	Pyrometry
Rudolph Technologies, Inc.	http://www.rudolphtech.com/home	Ellipsometer
SemiTest	http://www.semitest.com	Epi layer resistivity
Scientific Systems	http://www.scientificsystems.physics.dcu.ie	Langmuir probe, plasma impedance
SC Technology	http://www.sctec.com/sctinfo.htm	Reflectometry for coaters
Sigma Instruments	http://www.sig-inst.com	QCM
Sofie Instruments	http://www.microphotonics.com/sofie.html	Interferometer, optical emission spectroscopy (OES), ellipsometer
Sopra	http://www.sopra-sa.com	Spectroscopic ellipsometry
Spectra International	http://spectra-rga.com	RGA
Spectral Instruments	http://www.specinst.com	Spectrophotometers, CCD cameras
Therma-Wave	http://www.thermawave.com	Thin film and implant metrology
Thermionics	http://www.thermionics.com	Diffuse reflectance spectroscopy
Verity Instruments	http://www.verityinst.com	OES, spectral ellipsometer

cost, reliability, ease of integration into the processing tool, with sensitivity to equipment, and process variations over a broad range of processing conditions. The highest level sorting will be by the major process state (temperature, gas phase composition, plasma properties, etc.) and wafer-state (film thickness, thickness uniformity, resist thickness and profile, etc.) sensors. The major focus is on the technology behind each sensor. Applications will be described only when they are not necessarily obvious from the nature of the sensor. Any particular application example is not intended to promote that particular brand of sensor, but (1) it may be the only available sensor based on that technology or (2) the specifics may be required to provide a proper explanation for the use of that type of sensor. Links to the major companies selling sensors are in Table 25.1.

25.2 Process State Sensors

Sensors exist for monitoring both the process state of a particular tool and the wafer state of the processed wafer. The wafer state is of course, the critical parameter to be controlled, hence measurement of the appropriate wafer-state property is clearly the most effective means for monitoring and controlling a manufacturing process. However, this is not always possible due to

- lack of an appropriate sensor (technology limitation);
- lack of integration of appropriate sensors into processing tools (cost, reliability limitations).

In the above cases, the alternative is to monitor the process state of the manufacturing tool. In many cases, this is an easier task achieved with less expensive sensors. Nonintrusive RF sensors can be connected to the RF input lines, or the tuner, of a RF-powered processing tool. A range of optical techniques exists which require only an optical access to the processing chamber. Historically, the most predominant use of such process state sensors has been for endpoint determination. This is generally performed by the continuous measurement of an appropriate signal (e.g., intensity at a specific wavelength) during the processing of a wafer, looking for a change in the magnitude of the signal. Aside from endpoint determination, the availability of process state sensors in OEM processing tools is generally paced by integration issues (electrical, optical, cost, reliability). In addition, there is generally a lack of the necessary models that relate these process state measurements to the critical wafer-state properties. Especially due to this limitation, many of the process state sensors are typically employed for fault detection. This is the simplest use of such sensors, where the output is monitored in a univariate or multivariate statistical method, to determine deviations from the "normal" processing conditions. This methodology has a significant payback to manufacturing yield by an early determination of operational anomalies and hence the decreased misprocessing of wafers. Using process state sensors for process control requires much more rigorous models between the sensor signal(s) and the wafer-state parameter. The following is a description of the sensors that have, or soon will be, integrated into OEM processing tools for use in FDC or MBPC.

25.2.1 Temperature

The measurement and control of wafer temperature and its uniformity across the wafer are critical in a number of processing tools, such as rapid thermal processing (RTP), chemical rapor deposition (CVD), physical rapor deposition (PVD), and epitaxial (EPI) used for film growth and annealing. Historically, the most commonly used temperature measurement techniques are thermocouples and infrared pyrometry. Infrared pyrometry is based on analysis of the optical emission from a hot surface. It is dependent on two main variables: field of view of the detector and the optical properties of the material, such as refractive indices and emissivity. While useful only above 450°C due to the low emissivity of semiconductors in infrared, pyrometry has been commercialized and is widely utilized in semi conductor (SC) manufacturing tools. A newer technique is diffuse reflection spectroscopy (DRS), which provides noncontact, in-situ optical method for determining the temperature of semiconducting

TABLE 25.2 Pro and Con for Four Temperature Measurement Techniques

Technique	Advantages	Disadvantages
Thermocouple	Easy to use Low cost	Cannot be used in hostile environment Requires mechanical contact with sample Sensitivity depends on placement
Pyrometer	All-optical, noninvasive Requires a single optical port	Unknown or variable wafer backside emissivity Limited temperature range Sensitive to all sources of light in environment
Diffuse reflectance spectroscopy	Optical, noninvasive Directly measures substrate temperature Insensitive to background radiation Can be applied to a wide range of optical access geometries Wafer temperature mapping capability	Requires two optical ports Relatively weak signal level
Acoustic thermometry	Sample emissivity not a factor Wide temperature range	Intrusive to the reaction chamber Physical contact required

Source: Adapted from input by Booth, J., Thermionics Northwest, Port Townsend, WA. http://www.thermionics.com

substrates. The technique is based on the optical properties of semiconductors; specifically that the absorption coefficient rapidly increases for photon energies near the band gap of the material. Hence, a semiconducting wafer goes from being opaque to transparent in a spectral region corresponding to its band gap energy. A temperature change of the semiconductor is accompanied by a change in the band gap, which is then reflected as a shift of this absorption edge. Recently, an acoustic technology has been developed. The advantages and disadvantages of these four techniques are presented in Table 25.2.

Thermocouples are sometimes used for temperature measurement in processing tools. Since they have to be located remotely from the wafer, temperature errors of more than 100°C are possible; with no means for monitoring the temperature distribution across the wafer. Hence, this sensor is not widely used in SC manufacturing tools, so it will be omitted from this discussion.

25.2.1.1 Pyrometry

Precise wafer temperature measurement and tight temperature control during processing continue to be required because temperature is the most important process parameter for most deposition and annealing processes performed at elevated temperature [2]. As device features become smaller, tighter control of thermal conditions is required for successful device fabrication.

Optical temperature measurement is historically the primary method for in-situ wafer temperature sensing. Known as pyrometry, optical fiber thermometry, or radiation thermometry, it uses the wafer's thermal emission to determine the temperature. The optical fibers (sapphire and quartz), or a lens, are mounted on an optically transparent window on the processing tool and collects the emitted light from, in most cases, the backside of the wafer. The collected light is then directed to a photo detector where the light is converted into an electrical signal.

25.2.1.1.1 Theory of Operation

All pyrometric measurements are based on the Planck equation written in 1900, which describes a blackbody emitter. This equation basically expresses the fact that if the amount of light emitted is known and measured at a given wavelength, then the temperature can be calculated.

As a consequence of this phenomenon, all pyrometers are made of the following four basic components:

- collection optics for the emitted radiation;
- light detector;

- amplifiers;
- signal processing.

There are thousands of pyrometer designs and patents. A thorough description of the theory and the many designs, as well as the most recent changes in this field are well summarized in recent books and publications [3–6].

The two largest problems and limitations with most pyrometric measurements are the unknown emissivity of the sample—which must be known to account for the deviations from black-body behavior, and stray background light. In addition, the measurement suffers from a number of potential errors from a variety of sources. While the errors are often small, they are interactive and vary with time. The following is a summary of these sources of error, roughly in order of importance:

1. *Wafer emissivity*: worst case is coated backsides with the wafer supported on pins.
2. *Background light*: worst case is RTP and high energy plasma reactors.
3. *Wafer transmission*: worst at low temperatures and longer wavelengths.
4. *Calibration*: it has to be done reproducibly and to a traceable standard.
5. *Access to the wafer*: retrofits are difficult for integrating the sensor into the chamber.

The following problems will become much more important as the previous problems are minimized by chamber and pyrometer design.

6. *Pyrometer detector drift*: electronics (amplifiers) and photo detectors drift over time.
7. *Dirt on collection optics*: deposition and outgassing coat the fiber or lens.
8. *Changes in alignment*: moving the sensor slightly can cause an error by looking at a different place on the wafer or by changing the effective emissivity of the environment.
9. *Changes in the view angle*: changes the effective emissivity and hence the measured temperature.
10. *Changes in wavelength selective filters*: oxidation over years will change the filters.

Careful design of the *entire pyrometer environment*, not just the pyrometer itself, can serve to minimize these problems. Nearly all single-wafer equipment manufactures now have pyrometry options for their tools. Properly designed sensor systems in single-wafer tools can accurately measure wafers quickly (one-tenth of a second) from about 250 to 1200°C with resolution to 0.05°C.

The continually increasing process requirements will continue to push the pyrometric measurement limits to lower temperatures for plasma assisted CVD, cobalt silicide RTP, and plasma etch. The need for lower temperatures will mandate more efficient fiber optics. There will be a continued improvement in real-time emissivity measurement. Physical properties other than thermal radiance will also be used for measuring wafer temperatures. Repeatability will have to improve. The problem of unknown or changing emissivity is being addressed with the implementation of the ripple technique [7], which takes advantage of the modulation in background light to measure real-time wafer emissivity. This method can measure emissivity to ± 0.001 at a rate of 20 times per second.

25.2.1.2 Diffuse Reflectance Spectroscopy

25.2.1.2.1 Theory of Operation

Semiconductor physics provides a method for the direct measurement of substrate temperature, based on the principle that the band gap in semiconductors is temperature dependent [8]. This dependence can be described by a Varshni equation [9]

$$E_g(T) = E_g(T = 0) - \frac{\alpha T^2}{(\beta + T)} \tag{25.1}$$

where α and β are empirically determined constants. The behavior of the band gap is reflected in the absorption properties of the material. If a wafer is illuminated with a broadband light source, photons with energy greater than the band gap energy are absorbed as they pass through the material. The wafer is transparent to lower energy (longer wavelength) photons. The transition region where the material goes

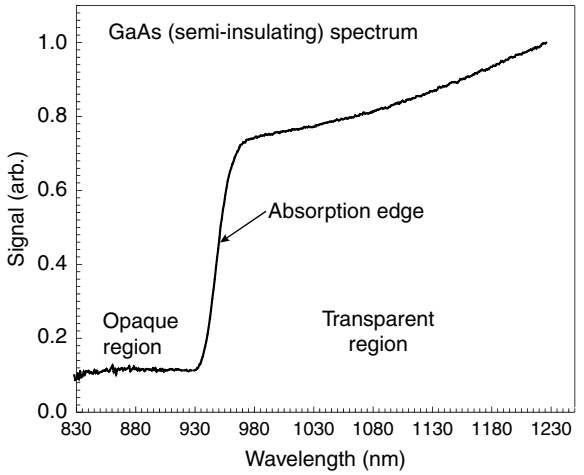


FIGURE 25.2 Spectrum of 350-mm thick gallium arsenide (GaAs) wafer showing the absorption edge where the wafer goes from being opaque to transparent. (Adapted from input by Booth, J., Thermionics Northwest, Port Townsend, WA. http://www.thermionics.com)

from being transparent to opaque occurs over a relatively narrow energy (or wavelength) range. Thus, a plot of light signal level passing through the material as a function of wavelength or energy (its spectrum) yields a step-like absorption edge, as shown in Figure 25.2. The position of the absorption edge shifts in energy or wavelength as the substrate temperature changes. The magnitude of the shift is material dependent.

25.2.1.2.2 Band Edge Measurement Techniques

There are several different techniques for measuring the optical absorption of semiconductors in-situ: transmission, specular reflection, and diffuse reflection.

Transmission measurements observe light that passes through the substrate. This technique has the advantage of providing a high signal-to-noise ratio (SNR). The main drawbacks are that to implement this approach one must have optical access to both sides of the sample and one is limited to single point temperature monitoring.

Reflection measurements from a sample can be separated into two components: specular reflection and diffuse reflection. Specularly reflected light is the familiar "angle of incidence equals angle of reflection" component. The diffusely scattered portion is light that is reflected from the sample over 2π steradians and carries with it color and texture information about the sample. The main differences between diffuse and specular reflection are:

1. Diffuse reflection can be seen over a wide range of angles, while specular reflection can only be observed from a narrow region in space.
2. Diffuse reflection is much weaker than specular reflection.
3. Both diffuse and specular reflections carry with them color and texture information. However, it is much easier to read this information from diffuse reflection because it is not buried under a large background signal level.

25.2.1.2.3 Band Edge Thermometry Limitations

There are two limitations to band-edge temperature measurement techniques: (1) free carrier absorption in the material and (2) coating the substrates with opaque material.

In semiconductors, there are a number of different phenomena that contribute to light absorption in the material. One of these is the so-called free carrier absorption. This absorption term is related to the number of carriers excited into the conduction band from the valence band. Free carriers are thermally excited and the free carrier absorption increases with sample temperature. This absorption occurs over a broad wavelength range and the material becomes more opaque as the temperature rises. The substrate band edge feature decreases in intensity until, above a material dependent threshold temperature, it can no longer be seen. In general, the smaller the band gap, the lower the temperature at which the substrate will become opaque. For silicon, the upper temperature limit for DRS temperature measurement is approximately 600°C, while for gallium arsenide the upper limit is estimated to be above 800°C.

The second limitation arises when the substrate is covered with an opaque material such as a metal. In this situation, light cannot penetrate the metal layer and consequently no band edge spectra can be recovered. Two remedies for this situation are: leaving an open area on the substrate for temperature monitoring purposes and viewing the substrate from the nonmetallized side.

25.2.1.2.4 Temperature Measurement: Practice and Applications

The initial application of band edge thermometry [10] measured the heater radiation that passed through a substrate to deduce the sample temperature. This approach suffered from a decreasing SNR as the heater temperature dropped. As in pyrometry, it was susceptible to contamination from other light sources present in the processing environment. Later on [11–13], two key innovations were introduced to the original transmission technique: (1) the use of an external modulated light source to illuminate the sample and (2) collection of the diffusely reflected source light to measure the band edge. The DRS technique is shown schematically in Figure 25.3. Part of the light is specularly reflected, while the other is transmitted through the sample. At the back surface, some of the transmitted light is diffusely scattered back towards the front of the sample. The collection optics are placed in a nonspecular position and the captured diffusely scattered light is then analyzed. The net result is a transmission-type measurement from only one side of the wafer.

A DRS temperature monitor [14] is shown schematically in Figure 25.4. The system consists of a main DRS module housing the monochromator, power supplies, and lock-in amplifier electronics. The light source and collection optics attach to the exterior of the processing chamber. Data collection and analysis are controlled by a stand-alone computer. The system can output analog voltages for direct substrate temperature control and can communicate with other devices using an RS-232 interface. The device is capable of 1 s updates with point to point temperature reproducibility of ± 0.2°C. The system can read

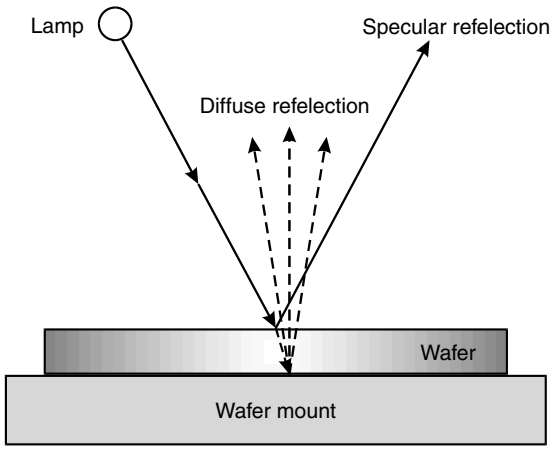


FIGURE 25.3 Schematic showing the diffuse reflection spectroscopy measurement technique. (Adapted from input by Booth, J., Thermionics Northwest, Port Townsend, WA. http://www.thermionics.com)

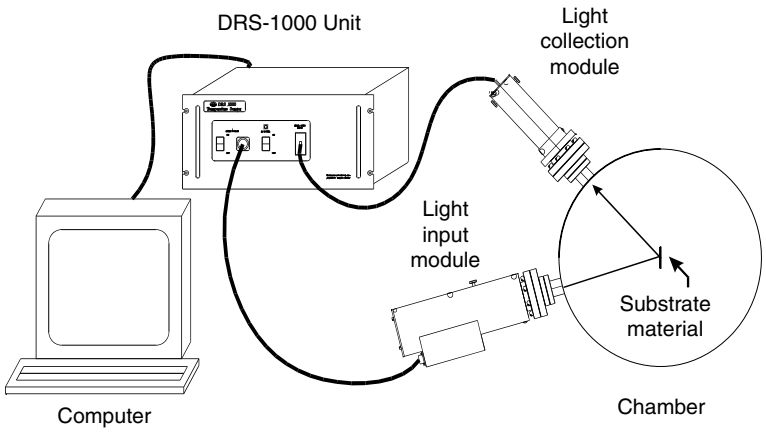


FIGURE 25.4 Diffuse reflection spectroscopy 1000™ temperature monitoring system schematic. (Adapted from input by Booth, J., Thermionics Northwest, Port Townsend, WA. http://www.thermionics.com)

silicon, gallium arsenide, and indium phosphide substrates from below room temperature to 600, >800, >700°C, respectively.

The DRS technology has been applied to both compound semiconductor and silicon processing. In molecular beam epitaxy and its related technologies, such as chemical beam epitaxy (CBE), material is grown layer by layer by opening shutters to molecular sources. The quality of the layers depends, in part, on the temperature and temperature uniformity of the substrate. In typical growth environments, the wafer temperature is controlled by a combination of thermocouple readings and pyrometer readings. The DRS can monitor and control the temperature of a gallium arsenide (GaAs) wafer in a CBE tool to well within ± 1°C [15]. Even though band edge thermometry has a fixed upper temperature limit of ~ 600°C for silicon, this technique is still applicable to several silicon processing steps, such as silicon etching [16], wafer cleaning, and wafer ashing.

25.2.1.3 Acoustic Wafer Temperature Sensor

Recently, a new technology has been developed for real-time wafer temperature measurement in semiconductor processing tools [17]. This product [18] is based on state-of-the-art acoustic thermometry technologies developed at Stanford University. This sensor fills the need for real-time wafer temperature measurement independent of wafer emissivity, especially in the sub-600°C process regime. It is compatible with plasma processes and mechanical or electrostatic wafer clamping arrangements.

25.2.1.3.1 Theory of Operation

The velocity of acoustic waves in silicon is a very linear function of temperature. Acoustic thermometry accurately measures velocity of an acoustic wave on the silicon wafer to determine the temperature of the wafer. The acoustic thermometer determines the velocity by very accurately measuring a delay between two points at a known distance.

In its simplest implementation, the acoustic thermometer contacts the wafer with two pins, as shown in Figure 25.5. One pin is a transmitter, and the other a receiver. Both pins have a piezoelectric transducer mounted to their lower end. Both piezoelectric transducers can turn an electrical excitation into an acoustic excitation, or vice versa. The tops of the two pins touch the wafer, to allow the transfer of acoustic energy between the wafer and the pins. The pins can be of any inert, relatively stiff material, such as quartz (fused silica) or alumina.

An electrical excitation pulse excites the transmitter pin's transducer to initiate the measurement. This excites an acoustic wave that propagates up the transmitter pin. When the wave reaches the top

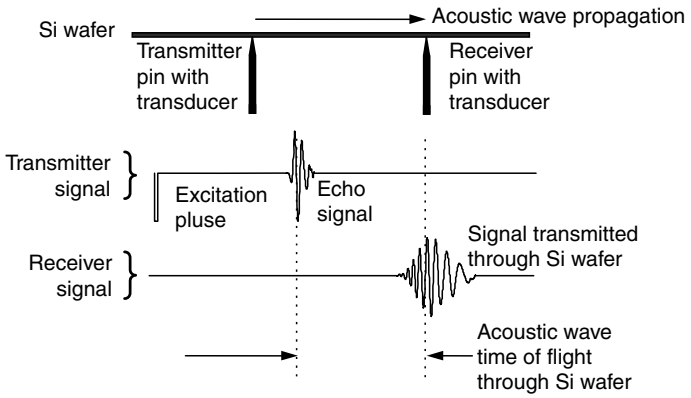


FIGURE 25.5 Geometry and electrical signal for acoustic thermometry. (Adapted from input by Sensys Instruments, Sunnyvale, CA.)

of the pin, two things happen to the acoustic energy. Most of the energy reflects back down the pin. The reflected energy gives rise to an electrical echo signal at the transmitter's transducer. A small amount of the acoustic energy enters the silicon as an acoustic wave, which propagates out from the transmitter pin. When the acoustic wave reaches the receiver pin, a small portion of its energy excites an acoustic wave that propagates down that pin. That wave produces the electrical signal at the receiver pin's transducer. The echo signal corresponds to propagation up a pin and down a pin. The received signal corresponds to propagation up a pin, across the wafer, and down a pin. If the pins are identical, the delay between the received and the echo signals corresponds to propagation on the wafer over the known distance between the transmitter and the receiver pins. The ratio of the distance to the delay is the velocity of the acoustic wave in the wafer, and is thus an indicator of the wafer temperature.

The temperature sensor is the wafer itself. Since the acoustic thermometer measures the velocity of propagation between the two pins, it measures an average temperature of the wafer along the propagation path between the pins, and not just the temperatures at the pins. Additional pins can yield average temperatures over multiple zones on the wafer. The acoustic wave extends across the whole thickness of the wafer. Thus, the measured temperature is an average over the thickness of the wafer as well. It is not particularly sensitive to the lower surface of the wafer, even though the pins only touch the lower surface.

This sensor can be designed to operate as the lift pin mechanism in single-wafer processing tools. The sensor does not influence or contaminate the wafer environment. The pin assembly has to be integrated into the OEM tool for the successful deployment of this technology.

25.2.2 Gas Phase Reactant Concentration

Most semiconductor manufacturing processes are chemical in nature. Typically, there is a chemical reaction between a gas phase chemical species (or mixture) with the surface layer of the silicon wafer. These reactions can be kinetically controlled, hence the interest in the wafer temperature. But they can also be controlled by the composition and concentration of the gas phase reactants. These parameters therefore have to be monitored and controlled in order to provide consistent, reproducible reactions, and wafer-state properties. Processing tools always control the primary components of these gas phase mixtures (flow rate of the gases, pressure). However, there is still a clear need for measuring the composition and/or the concentration of individual species to detect:

- changes in the chemistry as a function of the chemical reaction with the wafer, i.e., endpoint for etch;
- changes in the chemistry due to spurious conditions or faults (e.g., leaks, wrong gas);
- rapid changes or a slow drift in the gas phase chemical composition due to reactor chamber effects, such as cleaning, residue formation on the walls, wear of consumable parts within the reactor.

A large group of in-situ sensors for gas phase monitoring are spectroscopic in nature, as these are truly nonintrusive sensors. They require optical access through a single (sometimes opposing) windows, which are made of materials that can provide a vacuum seal and are transparent to the wavelengths being employed. The sensors analyze the composition of the gas phase via absorption or emission methods, as appropriate for a given process. The spectral range is from the UV through the IR, depending upon the nature of the information required. Mass spectroscopy, in the commonly used residual gas analysis (RGA) mode, provides another class of in-situ sensors used for monitoring gas composition. These are based on the analysis of the masses of the species (specifically, m/e) entrained in the gas flow. Sampling can be performed via a pin-hole orifice to the processing chamber, or by sampling the effluent from the reactor. A typical installation requires differential pumping of the RGA, although some of the recent systems do not have this requirement in selected low-pressure applications.

25.2.2.1 Optical Emission Spectroscopy

Optical emission spectroscopy (OES) is based on [19] monitoring the light emitted from a plasma during wafer processing and is used to gain information about the state of the tool and the process. It exploits the fact that an excited plasma emits light at discrete wavelengths which are characteristic of the chemical species present in the plasma. The intensity of the light at a particular wavelength is generally proportional to both the concentration of the associated chemical species and the degree of plasma excitation.

An OES system consists of a viewport to the plasma chamber, an optical coupling system, an optical detector incorporating some means of isolating the wavelength of interest, and a computer or processor to acquire and analyze the spectral image. The viewport is either a window in the reactor or a direct optical feedthrough into the chamber. The OES requires a direct view of the portion of the plasma immediately above the wafer, but not the wafer itself, so the placement of the viewport is not too restrictive. If ultraviolet wavelengths are to be monitored, the window must be of fused silica and not ordinary glass. A number of OES sensor systems are commercially available [20,21] and most OEM plasma tools come with their own on-board OES systems. The typical configuration is shown in Figure 25.6. The optical components and the other associated concerns with OES systems are described in the next sections.

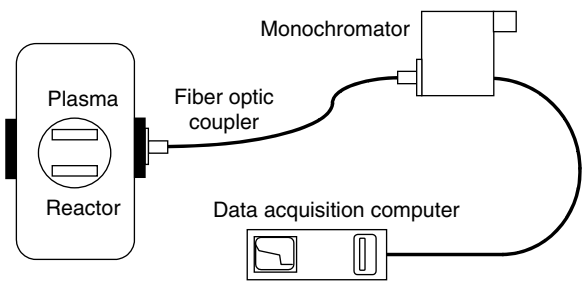


FIGURE 25.6 Optical emission sensor configuration. (Adapted from input by Whelan, M., Verity Instruments, Carrollton, TX. http://www.verityinst.com)

25.2.2.1.1 Fixed-Wavelength Systems

There are several types of optical detectors for OES. Simple systems use fixed bandpass filters for wavelength discrimination. These are stacks of dielectric films, and have a bandpass of typically 1–10 nm and a peak transmission of about 50%. The light that is passed by the filter is converted to an electrical signal either by a photodiode or by a photomultiplier tube (PMT). The advantages of these systems are low cost and high optical throughput; disadvantages are the limited spectral information and the mechanical complexity involved in changing the wavelength being monitored.

25.2.2.1.2 Monochromators and Spectrographs

More flexibility is afforded by systems which incorporate a monochromator. A monochromator consists of a narrow entrance slit, a diffraction grating, and an exit slit. Light falling on the grating is dispersed into a spectrum, the diffraction angle being dependent upon wavelength. The light is reimaged onto another slit, which provides wavelength discrimination. By turning the grating, any wavelength within the range of the instrument can be selected. Changing the width of the entrance and exit slits changes the bandpass of the system. Automated systems in which the wavelength can be altered automatically under computer control are often used. The throughput of a monochromator is much lower than that of a bandpass filter, hence PMTs are normally used for light detection in these systems.

A variant of the monochromator is the spectrograph. It uses a fixed grating, and a solid-state detector array instead of an exit slit and PMT. The advantage of a spectrograph over a monochromator is that many wavelengths can be monitored at once. This is significant for situations where information has to be based on an entire spectral scan, not from only a single spectral peak (an example is in Section 25.2.2.1.6.2).

In the typical installation, light is transferred from the viewport on the tool to the detector head via an optical coupler. Some fixed-wavelength devices mount directly onto the chamber, and no coupler is needed. In other cases, all that is required is to mount the detector against the window on the chamber. More typically, however, space or field-of-view considerations require the use of an optical coupling system. Optical fiber bundles are typically used to bring light into monochromators and spectrographs. This permits locating the detector at a convenient place away from the reactor. Attention to details such as f-number matching is required to prevent unnecessary loss of sensitivity. Suppliers of OES equipment can generally provide couplers which are matched to their equipment.

In the typical installation in a semiconductor production environment, an OES system usually consists of a detector head interfaced to a PC running a Windows or NT environment. The computer contains an A/D card for digitization of the analog signal from the detector, and runs an application that performs various functions: processing, data display, endpoint detection, and communication with the process controller.

25.2.2.1.3 Charge-Coupled Device Array Spectrometers

A charge-coupled device (CCD) array spectrometer consists of a slit, dispersion grating, and a CCD array detector. Light is typically delivered to the slit by a subminiature (SMA) connected fiber optic cable between the spectrometer and a fixture on the window of the reactor. A large number of CCD array spectrometers exist in the commercial market. They can be sorted into two main categories based on the characteristics of the CCD array and the form-factor of the final product.

The first category is characterized [22] by the "spectrometer-on-a-card" format, primarily aimed as a portable diagnostic tool for use with existing PCs. At least two such systems are currently available [23,24]. This means the spectrometer, say with a 2048 pixel CCD array, and all associated electronics can fit inside the PC, plugging into an expansion slot with an external fiber optic connection to the light source. This has some obvious benefits:

- Compact size: A substantial benefit in the high-cost clean room space.
- Portability: It can be used as a roaming diagnostic tool.

- Multiple spectrometers on a single card: Simultaneous spectral analysis in multichamber tools, or spatially resolved analysis [85] in a single chamber.
- Low cost: Cost is not a barrier to the use of this tool.

While there are benefits to these systems, there are some limitations that need to be understood. Specifically, the CCD array typically used in these systems has four inherent deficiencies (relative to PMTs), which if not adequately offset, will prevent the CCD array from surpassing or even matching PMT system performance. The four inherent CCD array deficiencies (relative to PMTs) are:

1. small detector pixel height (limited aspect ratio, e.g., 1:1), which limits the sensitivity of certain CCD array devices,
2. absence of inherent CCD array device "gain" (unity gain), which further limits the sensitivity,
3. poor UV response of certain CCD array devices, and
4. limited spectral resolution, due to typical CCD array configuration into very short focal length optical spectrographs, which exhibit more limited wavelength dispersion than is sometimes the case for PMTs, and which generally also exhibit severe, uncompensated, internal spectral imaging aberrations normally inherent with very short focal length optical spectrograph designs.

Fortunately, solutions exist that provide offsetting factors for each of the above-mentioned CCD array deficiencies. A brief description of these solutions provides background for understanding the key characteristics of CCD array spectrometers.

Concerning the problem of "small detector pixel height," offsetting factors include greatly reduced CCD dark current and noise (especially with small pixel areas), availability of selected array devices having greater than 1:1 (*h*:*w*) pixel aspect ratio (e.g., 20:1), and availability of 1D (vertical), internal, secondary, light-concentrating optics with certain CCD array spectrographs. A relatively "tall" spectral image is thereby height-focused onto a much shorter array pixel, thus concentrating the light and increasing the signal, without increasing the dark current or affecting spectral resolution (image width).

Concerning the problem of "absence of inherent CCD array device gain (unity gain)" relative to high gain PMTs, offsetting CCD array sensitivity factors include the natural integrating properties of array pixels, and an inherent CCD array quantum efficiency, which far exceeds that of PMTs.

Collectively, these offsetting factors are so effective that CCD arrays can be rendered sufficiently sensitive to achieve a "full well" device charge count (saturation) for prominent spectral features within the range of 400 ms (or less) exposure time, even with the dimmest of plasma etching experiments. When the light level is quite high, CCD array exposure times may typically be as low as 10 ms, or even less. The high light level allows, for example, 20 or even 40 separate (10 ms) exposures to be digitally filtered and signal averaged (co-addition) for each of 2048 array pixels. Digitally filtering and signal averaging this, many exposures provide a major statistical enhancement of the SNR. In addition, data from several adjacent wavelength pixels may optionally be binned (software summation) in real time for even more SNR enhancement, in cases where spectral resolution is not critical.

Concerning the problem of "poor UV response," offsetting factors exist, in the form of fluorophore coatings applied directly to the detector pixels. Satisfactory UV response is thereby achieved.

The problem of "limited spectral resolution" is one of the most basic problems in using CCD array systems. At most, CCD arrays are only about 1-in. long (e.g., 21–28 mm). This means the entire spectrum must be compressed to fit the 28-mm array length, which limits the spectrographic wavelength dispersion that may be employed. There is an additional resolution and spectral range tradeoff in the choice of gratings. The total wavelength coverage interval of a CCD array is determined by array dimensions and the spectrograph focal length and grating ruling density, which together establish the wavelength dispersion. For an array of fixed dimensions, and a spectrograph of fixed focal length, coarsely ruled gratings (600 grooves/mm) provide less dispersion and hence lower resolution, but a larger total wavelength coverage interval. Finely ruled gratings (1200 or 2400 grooves/mm) provide more dispersion and higher resolution, but a smaller total wavelength coverage interval. Centering of a given wavelength range is specified by the user and is fixed at the factory by adjusting the grating angle.

The second category of spectrographs is characterized by high-performance CCD arrays, with applications aimed at stand-alone use (PC, or laptop, not necessarily required) or integration into OEM processing tools. These are based [19] on research-grade CCD spectrographs that are available with performance that equals or exceeds that of PMT-based systems. For maximum sensitivity, they employ cooled, back-illuminated CCD area arrays. The CCD is operated in a line-binning mode, so that light from the entire vertical extent of the slit is collected. These devices have peak quantum efficiencies greater than 90%, and over 40% throughout the typical spectral range of interest (200–950 nm), when compared with a peak value of 20% typical of a PMT. This means that about one photoelectron is created for every two photons to reach the detector. The best such devices have readout noise of only a few electrons, so that the signal-to-noise performance approaches the theoretical limit determined by the photon shot noise. However, the traditional size, cost, and complexity of these instruments make them impractical for use for routine monitoring and process control.

Nonetheless, many of these features are beginning to appear in instruments priced near, or even below $10K. Spectrographs in this price range are available [20,25,26], which employ a cooled, back-illuminated CCD with a 3-mm slit height, and whose sensitivity matches or exceeds that of currently available PMT-based monochromators. If cost is the prevailing factor, lower cost can be achieved using front-illuminated CCDs. Performance suffers, since the quantum efficiency is reduced by a factor of 2, and the spectral response of these devices cuts off below 400 nm. Nonetheless, this is a cost-effective approach for less-demanding applications.

The issues of size and complexity are being addressed as well. One approach is to integrate the optical head together with the data acquisition and process-control functions into a single unit [20] with a small footprint and an on-board digital signal processor (DSP) for data analysis. Such a system can be integrated with the host computer for OEM applications, or be connected to a laptop for roaming applications.

Another advantage of such high-performance arrays is that they can be utilized as an imaging spectrograph, where the entrance slit is divided into multiple sections that can couple to different chambers. The resulting spectra can be read independently from the CCD. In this way, multiple spectra can be run on the same instrument.

25.2.2.1.4 Calibration, Interface, and Upkeep Issues

Implementing an OES analysis for a new process requires some expertise on the part of the process engineer. First, the spectral line or lines to be monitored must be chosen based upon a fundamental understanding of the spectral signature. Routine acquisition and signal analysis can then be performed by the sensor. The practical issue of the etching of the window or optical feed-through (or deposition on these components) has to be handled by cleaning or replacing these components. Some OEM vendors address these issues by heating, or recessing, the windows (to slow down deposition), or the installation of a honey-comb-like structure over the window (to cut down the deposition or etch on the major cross-sectional area of the window).

25.2.2.1.5 Routine Application—End Point Detection

By far the most widespread use of OES sensors is for endpoint detection. Historically, such sensors have been routinely used in plasma etch reactors for decades, since the process state (plasma) is a rich source of useful information. The fundamental principle behind endpoint detection is that as the etch proceeds from one layer (the primary layer being etched) to the underlying layer (the substrate), the gas phase composition of the plasma changes. For example, when etching a typical TiN/Al/TiN stack on an oxide substrate with a Cl-containing chemistry, there is a significant decrease in the AlCl product species with a corresponding increase in the Cl-reactant species, as the etch transitions from the bulk Al to the TiN and oxide layers. So a continuous monitoring of the 261-nm Al emission line intensity will show a decrease during the time when the Al film disappears. Traditional endpoint detection techniques have historically relied on numerical methods, such as threshold crossing, first derivative, or other combinatorial algorithms, which are manually devised to conform to the characteristics of a family of endpoint shapes

and can be tuned to declare endpoint for a typically anticipated signal change. Besides the endpoint indication—which is by far the most commonly generated information from this data—the slope of the endpoint signal (at the endpoint) can be used as an indicator of the nonuniformity of the etch process [27].

From the sensor point of view, endpoint detection has been well established for the last 15–20 years. The majority of OEM plasma etch tools have endpoint detection hardware integrated into the tool. Historically, these were simple, inexpensive photodetectors that viewed the plasma emission through an appropriately selected optical bandpass filter and an optically transparent window in the side of the reactor. In the newer generation tools, this optical signal is obtained by the use of short focal length grating monochromator that can be manually scanned to the correct wavelength for the specific process. These have the advantage of the readily variable wavelength selection, the higher spectral resolution required to optically separate closely overlapping spectral lines, and a higher sensitivity of the PMT detector (vs. the photodiode detectors).

25.2.2.1.6 Emerging Application—Endpoint Detection for Low Exposed Areas

For etch processes where the material etched is a significant percentage of the wafer surface area (e.g., metal etch), there is a large change in the plasma chemistry when this material is etched off, hence the endpoint signal is very strong and easily detected. The latest challenge in low-exposed area endpoint detection is for processes such as oxide etch in contact holes where the exposed area of oxide is less than 1% of the wafer area. This drives a very small chemistry change at endpoint, which in turn generates a very small change in the large emission signal intensity. The following sections provide details on the newly emerging software methods for detecting endpoint based on single-wavelength endpoint curves (Section 25.2.2.1.6.1) or on full spectral data (Section 25.2.2.1.6.2).

25.2.2.1.6.1 Neural Network Endpoint Detection

The shape of the typical single-wavelength endpoint curve is the natural by-product of the processing tool characteristics, the product design, and the influence of the tools and processes that precede the endpoint-defined tool step. As such, the endpoint shape of a given process exhibits a statistical variation derived from the numerous preceding processes that affect the state of the wafer supplied to the tool requiring endpoint control. It then becomes the challenge of the process engineer, working in conjunction with the endpoint controller system, to effect a practical control algorithm. This algorithm has to be unique enough to recognize the endpoint, but also general enough to comprehend the pattern variability, which is a consequence of the tool change (namely, window absorbance) and the product mix. This challenge can be imposing, requiring a lengthy empirical evaluation of numerous endpoint data files in an attempt to achieve the correct numerical recipe, which accurately and reliably declares endpoint for the full suite of endpoint pattern variations.

One approach [19] to this problem is a neural network-based endpoint detection algorithm [28]. It utilizes a fast training neural network pattern recognition scheme to determine the endpoint signature. Unlike traditional feed forward neural networks which require many pattern samples to build an effective network, the methodology employed with this approach minimizes the number of representative sample data files required for training to typically less than 10. The process engineer is not burdened with numerical recipe optimization. The following simple three-step procedure outlines the technique.

1. Acquire representative data files exhibiting a full range of endpoint patterns; new patterns can be later introduced into this data set.
2. Tag the endpoint patterns in the collected data files, i.e., identify the region in each data set that contains the endpoint.
3. Train the network—an automatic procedure completed in a few minutes.

This technology has been successfully demonstrated and used in etching oxide with as low as 0.1% open area. Ultimate limits for any specific application are tied to a number of variables. These include the type of tool, process, optical detector, and appropriate selection of emission wavelength(s) to monitor. As a caveat, it is important to note that this technique must "see" the evolution of a distinguishable pattern

in order to learn the shape and correctly identify its occurrence as endpoint. The shape may be subtle and complex, but it must be identifiable for successful results.

Another useful feature of this type of endpoint detector is its ability to recognize complex, unusual, nonmonotonic patterns. Since this technique employs a pattern recognizer, it is not limited by complex signal variations that can prove daunting for numerical recipes.

25.2.2.1.6.2 Evolving Window Factor Analysis of Full Spectra

With only one wavelength being monitored, random plasma fluctuations or detector/amplifier noise excursions can obscure the small intensity change that would otherwise serve to characterize endpoint in oxide etching of wafers exhibiting open areas of 1% or less. An alternate solution to this problem is to use the full spectrum available from the plasma. It is clear that significantly more useful data can be had if one measures the intensities in a broad spectral range vs. at a single wavelength of a chosen species. Since the spectral data still have to be obtained at a fast-enough rate to detect the endpoint, this drives the use of CCD array detectors. With the necessary dispersion optics, these CCD arrays can simultaneously measure the spectral intensities across a broad spectral range at a resolution determined by a number of factors described in Section 25.2.2.1.3. This sensor change generates a new dimension in the endpoint data, the spectral dimension (vs. a single wavelength). This, in turn, necessitates the development of algorithms that can make good use of this additional information.

In one particular case [23], an evolving window factor analysis (EWFA) algorithm is employed to obtain the endpoint information from the multivariate spectral data. An EWFA is a variant of the more classical evolving factor analysis (EFA) technique used for the analysis of ordered—in this case, by time—multivariate data. The EFA follows the singular values (factors) of a data matrix as new rows (samples) are added. In a manner similar to principal component analysis (PCA) analysis, EFA determines how many factors are included in the data matrix and then plots these against the ordered variable (time). Such algorithms are well defined and routinely available [29]. The EWFA variation is to consider a moving window of data samples (as in Figure 25.7), for computational ease.

The resulting data is a time-series plot of the appearance, or disappearance, of certain factors in the data set. So looking at the multiple spectral lines provide an increased endpoint signal sensitivity. A typical representation of this EWFA is in Figure 25.8, which shows two of the factors. Factor 3 (value 3) shows the temporal nature of the rotating magnetic field in this processing tool. Factor 4 (value 4) shows the endpoint signals from the etching of a four-layer oxide stack; the four endpoint signals are clearly identified in spite of the other temporal variations in the process (the rotating magnetic field). Automated endpoint detection in oxide etching has been shown to work with this technique down to 0.1% open area.

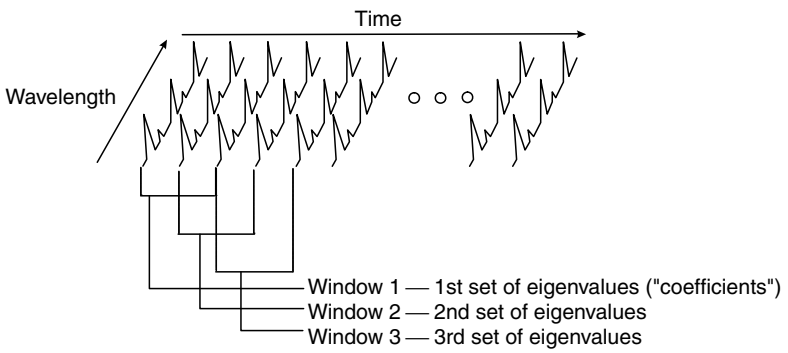

Time
Wavelength
Window 1 — 1st set of eigenvalues ("coefficients")
Window 2 — 2nd set of eigenvalues
Window 3 — 3rd set of eigenvalues

FIGURE 25.7 Evolving window factor analysis data matrix. (Courtesy of Bob Fry, Cetac Technologies, Inc., Omaha, NE.)

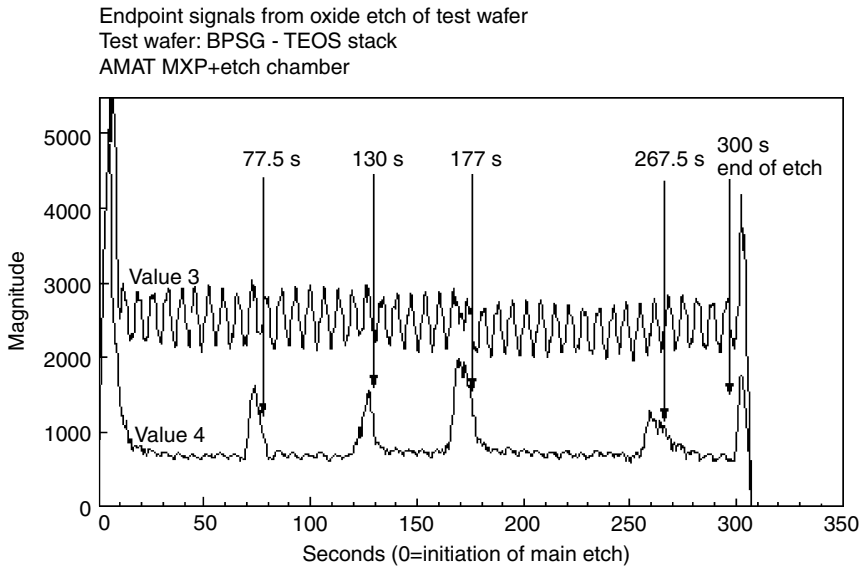


FIGURE 25.8 Evolving window factor analysis endpoint signal on "Value 4". (Courtesy of Bob Fry, Cetac Technologies, Inc., Omaha, NE.)

25.2.2.2 Fourier Transform Infrared

Infrared spectroscopy, in the mid-IR range of 1–20 μm, can provide a wealth of information about gas properties, including species temperature, composition, and concentration. Its application to gas phase analysis in SC manufacturing tools has been more limited than the use of visible spectroscopy for analyzing the optical emission of plasma processes. But there are some useful applications based on infrared spectroscopy, hence the technology will be described.

25.2.2.2.1 Theory of Operation

Complete spectra from 1.5 to 25 μm wavelength can be obtained in fractions of a second using a FTIR spectrometer. The core of a FTIR is typically a Michelson Interferometer consisting of a beam splitter and two mirrors, one of which moves [30]. As shown in Figure 25.9, incoming radiation in a parallel beam impinges on the beam splitter and is split roughly in half into beams directed at the mirrors. The reflected light recombines at the beamsplitter to form the outgoing radiation. If the mirrors are equidistant from the beam splitter, then the radiation recombines constructively. If the paths differ by one-fourth wavelength, then the beams combine destructively. As the moving mirror travels at constant velocity, the radiation is amplitude modulated, with each frequency being modulated at a unique frequency that is proportional to the velocity and inversely proportional to the wavelength. Thus, radiation with twice the wavelength is modulated at half the frequency. The key requirements for such a FTIR spectrometer are that they are vibration immune, rugged, permanently aligned, and thermally stable. Another key issue for accurate quantitative analysis is detector linearity. Mercury cadmium telluride (MCT) detectors are high sensitivity infrared detectors, but are notoriously nonlinear. Detector correction methods are required to linearize the response. All these requirements have been addressed, making FTIR a commercially available sensor [31] for possible use in SC manufacturing.

25.2.2.2.2 Exhaust Gas Monitoring Applications

Most of the sensors described in this chapter are used for sensing process or wafer-state properties during or after a given manufacturing process. However, FTIR lends itself well to another important aspect of SC manufacturing; fault detection based on exhaust gas monitoring from a reactor.

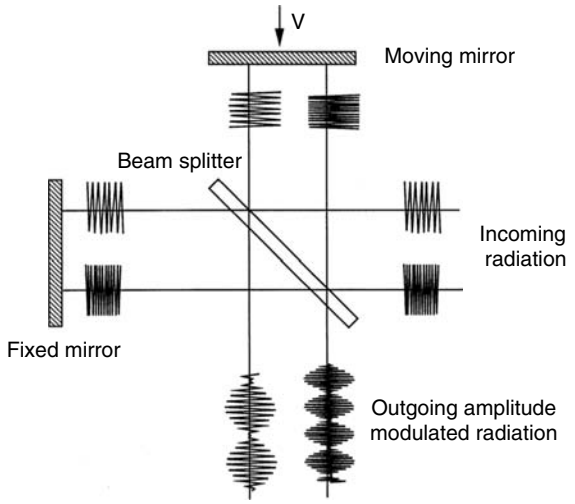


FIGURE 25.9 Modulation of radiation by a moving mirror interferometer. (Adapted from input by Peter Solomon, On-Line Technologies, East Hartford, CT.)

In one particular study [32] on a high-density plasma reactor, the spectrometer was positioned above the exhaust duct of the etch chamber. The IR beam was directed to a set of focusing and steering mirrors and into a multipasso mirror assembly enclosed in an exhaust line tee. This tee was placed between the turbo and the mechanical pumps. The multipass cell generated 20 passes through the tee to provide a 5-m path length. The exhaust gas passed through this in-line gas cell and spectra were collected at 1/cm resolution.

The data obtained in this study suggests that FTIR measurements can provide:

1. Exhaust gas monitoring, after the turbo-pump, providing a reproducible and rapid measurement of a rich variety of compounds produced during the wafer etch.
2. Identification of the mix of compounds which can be used to interpret an etching sequence, or the cleaning of a reactor by a reactive plasma.
3. Identification for the effects of incorrect chucking, incorrect plasma power, air leaks, low pressure gas feed.
4. Data for use in fault detection, for a reliable and automated fault detection and classification system.

FTIR can also be used for the analysis of the efficiency of large scale, volatile organic compound abatement systems.

25.2.2.3 Mass Spectroscopy/Residual Gas Analysis

In addition to the optical methods previously described, gases can also be analyzed by mass spectroscopy of the molecular species and their fragmented parts. The in-situ mass spectrometric sensor for gas analysis is commonly known as a residual gas analyzer.

25.2.2.3.1 Conventional Residual Gas Analyzer

Quadrupole devices are by far the most widely used in semiconductor manufacturing applications, typically for the maintenance and troubleshooting of process tools [33]. Leak checking, testing for gas contamination, or moisture and periodic RGA for tool qualification have been the main uses of RGA. In other applications, RGAs are used to establish a correlation between wafer's quality and measured contamination in the process chamber. Recently, RGAs have been used for in-situ monitoring to prevent

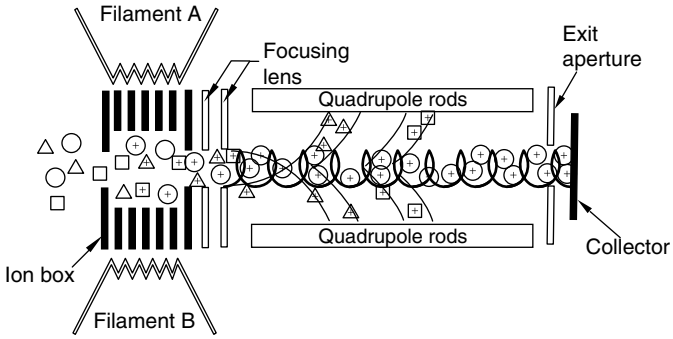


FIGURE 25.10 Quadrupole residual gas analyzer. (Adapted from input by Ferran, R.J. and Boumsellek, S., Ferran Scientific, Inc., San Diego, CA. http://www.ferran.com)

accidental scrap and reduce wafer-to-wafer variability. To be effective, RGAs have to be able to directly monitor both tool baseline pressures and process chemistries in a nonintrusive fashion.

25.2.2.3.1.1 Theory of Operation

Conventional RGAs operate by sampling the gases of interest through an orifice between the container for the gases (the processing chamber or exhaust duct in an SC manufacturing tool) and the residual gas analyzer (shown in Figure 25.10). In a conventional RGA, the pressure must be reduced below typical processing chamber pressures prior to ionization. This requires differential pumping and sampling of the process gases, making conventional RGAs a relatively bulky and expensive package. The following brief description of the three basic components of a quadrupole mass spectrometer analyzer—the ionizer, the mass filter, and the detector—are provided to facilitate the understanding of the sensors based on this technology.

25.2.2.3.1.2 Ionizer

Gas ionization is usually achieved using an electron-impact type process. Electrons are emitted from a hot filament (2200°C) using an electric current. Few metals have a low enough work function to supply currents in the milliampere range at such temperatures. Filaments are usually coated with materials with better thermo-emission properties. Typical coatings are thoria and yttria and typical base metals are tungsten, iridium, and rhenium. Electrons are then accelerated to acquire an energy in the 30–70 eV range, which corresponds to the highest ionization cross-sections for several gases. The ionization occurs in an enclosed area called the ion source. There are many types of sources, but the major distinction is between open and closed sources. The higher the pressure in the source, the greater is the sensitivity to minor constituents. The sensitivity is the minimum detectable pressure relative to the maximum number of ions produced in the source. A closed ion source has small apertures to introduce the sample gas from the process environment, to allow the electrons to enter the source, and to extract the ions into the mass filter. With the use of an auxiliary pump, the filaments and the mass filter and the detector are kept at a much lower pressure than the source. In addition to greater sensitivity, the advantages associated with closed sources are: (1) prolonging the filament lifetime in the presence of corrosive gases and (2) enabling electron multipliers to be used as ion detectors. However, the high complexity and cost associated with the aperture's precision alignment and the required high vacuum pump make closed-source-type instruments very expensive.

25.2.2.3.1.3 Mass Filter

The ions are extracted from the source and are focused into the entrance aperture of the mass filter with an energy, V_z. The mass filter is the cavity enclosed by the four parallel quadrupole rods arranged in a square configuration (see Figure 25.11). Typical diameter and length of the cylindrical rods are at

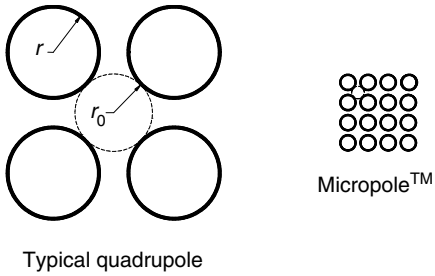


FIGURE 25.11 Array concept. (Adapted from input by Ferran, R.J. and Boumsellek, S., Ferran Scientific, Inc., San Diego, CA. http://www.ferran.com)

least 6 and 100 mm, respectively. The species moving through the filter are singly or multiply charged atoms or molecules. Filtering is the common term for selecting ions with a particular mass-to-charge ratio that possess a stable trajectory enabling it to reach the detector, while all other ions (with unstable trajectories) are filtered out. Filtering is accomplished by subjecting the ions to lateral forces generated by the combination of dc and RF voltages on the rods. The filtered mass and the mass resolution are given by

$$m = \frac{7 \times 10^6 V}{f^2 r_0^2} \tag{25.2}$$

$$\Delta m = \frac{4 \times 10^9 V_z}{f^2 l^2} \tag{25.3}$$

where V is the amplitude of the RF voltage, f is the RF frequency, r_0 is the radius of the inscribed circle, l is the length of the mass filter, and V_z is the ion energy.

25.2.2.3.1.4 Detector

The filtered ions are accelerated at an exit aperture to reach the detector. Two detection techniques are generally used: Faraday cups and electron multipliers. Faraday cups are in the shape of cavities in which collected ions and any secondary electrons are trapped to generate a current. The current is then converted to a voltage using a sensitive electrometer circuit. The limit of detection of these devices is gated by the ability to make more sensitive electrometers. Fundamental limitations associated with Johnson noise in resistors and the noise in the semiconductor junctions determine the lowest detectable current. Alternatively, there are techniques for multiplying the current in vacuum using a continuous dynode electron multiplier. This is shaped as a curved glass tube with the inside coating made of a high-resistivity surface ($PbO–Bi_2O_3$ glass) with a high secondary electron emission coefficient. A high voltage (3 kV typically) is applied between the ends of the tube. When filtered ions strike the active surface, a shower of electrons are produced and accelerated towards the opposite wall of the surface. Each electron leads to the emission of more electrons and the process is repeated along the length of the tube causing an avalanche of electrons. A multiplication or gain up to 10^7 can be achieved. However, the ability to emit electrons decreases with time. The time scale depends on the total number of electrons emitted, which in turn depends on the number of incident ions. At high pressures, large amounts of ions strike the surface causing a high rate of depletion and hence a shorter lifetime. Another important phenomenon related to the operation at high pressures is the "positive feedback." As the number of positive ions increases inside the tube, the gain can be drastically reduced since ions, accelerated in the opposite direction, interfere with the electron multiplication process. These phenomena limit the practical use of electron multipliers to the low-pressure ($<10^{-5}$ Torr) range.

25.2.2.3.2 Sensor-Type RGAs

25.2.2.3.2.1 Component Choices

A recent key development in RGA technology is the evolution of sensor-type RGA. These have miniaturized quadrupoles which allow mass-filter operation at nearly three orders of magnitude higher pressure, hence not requiring differential pumping of the sensor for many applications. The basis of these new systems is the substantially shortened quadrupoles, which provide a short path for ions to travel to the detector, hence minimizing the effect of collisional losses at the higher pressures. Their small size allows them to be mounted at several strategic and/or convenient locations without sacrificing any foot-print. This represents a major breakthrough with regards to the sensor size, cost, and ease-of-use. A number of sensor manufacturers provide such sensors [34,35]. But any miniaturization attempt has to be carried out without sacrificing mass spectrometry performances of conventional RGAs in term of sensitivity, mass resolution, and mass range.

The optimal use of these sensors requires an understanding of the interactions between the pressure range, the detection technique, and the required sensitivity. At low pressures (below 10^{-5} Torr), which conveniently coincides with the optimum operating pressure of the high-gain electron multipliers, high sensitivity to low partial pressure contaminants is readily achieved. This provides capability for sensitive determination of background moisture levels and low level leaks in vacuum system.

But with these smaller sensors currently available, the shorter path length of the ions allows RGA mass filters to operate at pressures in the milliTorr range, which also enables the direct monitoring of many semiconductor manufacturing processes. However, these pressures are too high for efficient operation of the electron multiplier detector. So one solution is to return to the use of a pressure throttling device (orifice) and a high vacuum pump. Aside from the cost and size penalties of this approach, there are more serious considerations that have to do with the retained gases, or lack thereof, on the analyzer chamber walls. This leads to measurements which do not necessarily reflect the gas composition of the process chamber, but reflect more on the state of the analyzer. This is very noticeable when the pressure is very low in the analyzer chamber. The lower the pressure in the analyzer chamber, the lower the required pressure of the high vacuum pump, since species not pumped will provide a background measurement which must be taken into account and may be variable with time and temperature. Another solution is to use a Faraday cup detector at these milliTorr pressures, but this sacrifices sensitivity due to the lower gain inherent in these detectors. The sensitivity is further decreased by the geometrical aspects of these miniature sensors since the reduction of the rod size, and hence of the diameter of the inscribed circle between the four rods, results in a smaller acceptance area for the ionized species to reach the detector.

A recent solution to this sensitivity issue at higher pressures has been the development of an array detector. In this configuration, an array of miniature quadrupoles compensates for the loss of sensitivity. The mass resolution and the mass range are maintained by increasing the RF frequency as seen in Equation 25.2 and Equation 25.3. While the array concept was introduced several decades ago, volume production has only recently been enabled by the new fabrication technologies used to handle microparts.

This sensor [36] comprises a 4×4 array of identical cylindrical rods (1-mm diameter) arranged in a grid-like pattern, where the cavities between the rods form a 3×3 array of miniature quadrupole mass spectrometers. The length of the rods is only 10 mm which enables the operation of the sensor at higher pressures (10 mTorr). It occupies less than 4 cm^3 total volume. The manufacturing method uses glass-to-metal technology to seal the rods and electrical pins. This technology provides lower manufacturing cost and physically identical sensors that are simple to calibrate. The replacement cost of these sensors is low enough to consider them to be consumables.

25.2.2.3.2.2 Calibration and Lifetime

The RGA sensor calibration is performed against capacitance manometers and need no field calibration other than for fault detection. The data are displayed in Torr or other acceptable pressure units and can be directly compared with the process pressure gauge. There are recommended practices published by the

American vacuum society (AVS) for calibrating the low-pressure devices, but the miniature high-pressure RGAs were developed subsequent to the practices being established.

At the high operating temperatures, RGA filaments react strongly with the ambient gases. This interaction leads to different failure mechanisms depending on the pressure and the chemical nature of these gases. Tungsten and rhenium filaments are volatile in oxidizing atmospheres (such as oxygen), while thoria and yttria-coated iridium filaments are volatile in reducing atmospheres (such as hydrogen). Corrosive gases, such as chlorine and fluorine, reduce filaments lifetime drastically. In fact, lifetime is inversely proportional to the pressure in the range of 10^{-5} to 10^{-2} Torr. This corrosion-limited lifetime favors systems that have readily and inexpensively replaceable detectors.

25.2.2.3.2.3 Sensor Interface to OEM Tools

Since both the process tools and the RGA sensors are vacuum devices, the pressure connections to the vacuum should be made with good vacuum practice in mind. Lower pressure operation increasingly requires the use of large diameter, short lines made with materials and processes, which provide low retentivity. Wherever possible, the source of the sensor should be in good pneumatic communication with the gases to be measured. Care should be taken to avoid condensation in the RGA of species from the process. This may require that the sensor and its mounting be heated externally. Temperatures as high as 150°C are common in these cases. Higher temperatures may require separating portions of the system from the sensor, which may deteriorate the system performance and add to the cost.

Quadrupole devices operate internally at very high RF voltages and therefore may radiate in the process chamber or to the ambient outside the chamber. Good grounding practices such as those used with process plasma RF generators are important.

Compared to other instruments on the process tool, RGAs generate large amounts of 3D data (mass, pressure, time) in a very short time. Invariably, the data from these devices are transmitted on command by a serial link. RS232, RS485, and proprietary protocols are all in use. Since multiple devices are commonly bused together, data systems such as a "Sensor Bus" will normally be used. Efficient use of the data by the tool controller represents a major challenge on the way to a full integration of these sensors into OEM tools.

25.2.2.4 Acoustic Composition Measurement

Although generally not well recognized, the composition of a known binary mixture of two gases can be determined by measuring the speed of sound in the mixture [37]. Very high sensitivity ($\sim$1 ppm) is available when a high molecular weight precursor is diluted in a low molecular weight carrier gas. Acoustic gas composition measurement is inherently stable and consequently accuracy is maintained over the long term. There are no components that wear and the energy levels imparted to the gasses are very low and do not induce any unintended reactions. These features make this technique ideal for many metal organic chemical vapor deposition (MOCVD) and CVD processes. This technique is not readily applicable if the individual gas species are unknown or if more than two species are present, because many combinations of different gas species may be blended to produce the same speed of sound. This lack of uniqueness does not pose a problem for blending gasses since a cascaded arrangement of sensors and controllers may be used to add one gas at a time. Each successive transducer will use the information for the blend from the previous instrument as one of its component gas' thermodynamic constants. The most obvious application is to determine the gas composition flowing through the tubes that supply mixed gasses to a reactor. However, there may be additional value to sampling gasses from within the reactor chamber or its outlet. This dual transducer arrangement can provide information on the efficiency, stability, and "health" of the ongoing process.

25.2.2.4.1 Theory of Operation

The speed of sound, C, in a pure gas is related to the gasses' fundamental thermodynamic properties as follows [38]

$$C = \sqrt{\gamma RT/M} \tag{25.4}$$

where γ is the specific heat ratio, C_p/C_v, R is the universal gas constant, T is the Kelvin temperature, and M is the molecular weight. The same equation form holds precisely for a mixture of gasses when appropriate values for γ and M are calculated based on the relative abundance of the individual species. Likewise, it is only an algebraic exercise to solve the resulting equation for the relative concentration of a mixture when the speed of sound is known or measured [39].

25.2.2.4.2 Sensor Configurations

Building a composition-measuring instrument using this fundamental thermal physics has been accomplished in two distinct ways. The first implementation measures the transit time for an ultrasonic ($\sim$15 kHz) pulse through the gas [40]. This time-of-flight implementation only requires a high resolution timer to measure the time between when a sound pulse is generated and its arrival at a receiver a distance, L, away. The second implementation measures the resonant frequency of a small chamber filled with the target gas mixture [39], as in Figure 25.12. All wetted components of this chamber are fabricated from high purity and electro-polished stainless steel and inconel. A precisely controlled frequency generator is used to stimulate the gas at one end of the chamber and the intensity of the transmitted sound is measured at the opposite end. Algorithms are designed to maintain the applied frequency at the gas chamber's resonance frequency. The chamber is precisely temperature controlled (to less than $\pm$0.03°C) and is carefully shaped to resonate in the fundamental mode at a low audio frequency (0.3–4.6 kHz). Low frequency operation allows the use of metal diaphragms which avoids process gas contact with the acoustic generators and receivers. Low frequency sound is also more efficiently propagated through the chamber than ultrasonic frequencies, resulting in useful operation at pressures as low as 70 Torr. Since the chamber's length is fixed and its temperature is carefully controlled, the speed of sound, C, is simply related to the resonant frequency, F, as $C = 2FL$, where L is the effective distance between the sending and the receiving elements. It is possible to resolve a gas filled chamber's resonant frequency, and therefore the speed of sound of the gas, to less than 1 part in 50,000 using the resonant technique. Even though the frequency generation method employed can generate frequencies only 0.1 Hz apart, even greater resolution may be achieved by measuring the amplitude at several frequencies around resonance and curve fitting the instrument's response to the theoretical shape of the resonance peak.

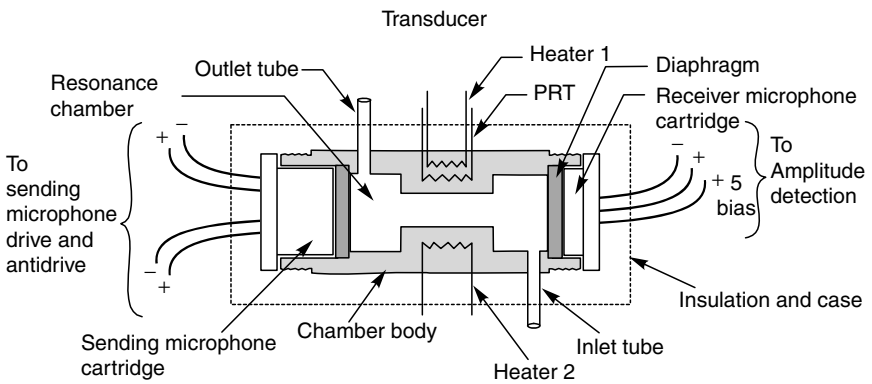


FIGURE 25.12 Cross-section of a transducer for a low frequency resonance type acoustic gas analyzer. (Adapted from input by Gogol, C.A., Leybold Inficon, East Syracuse, NY. http://www.inficon.com)

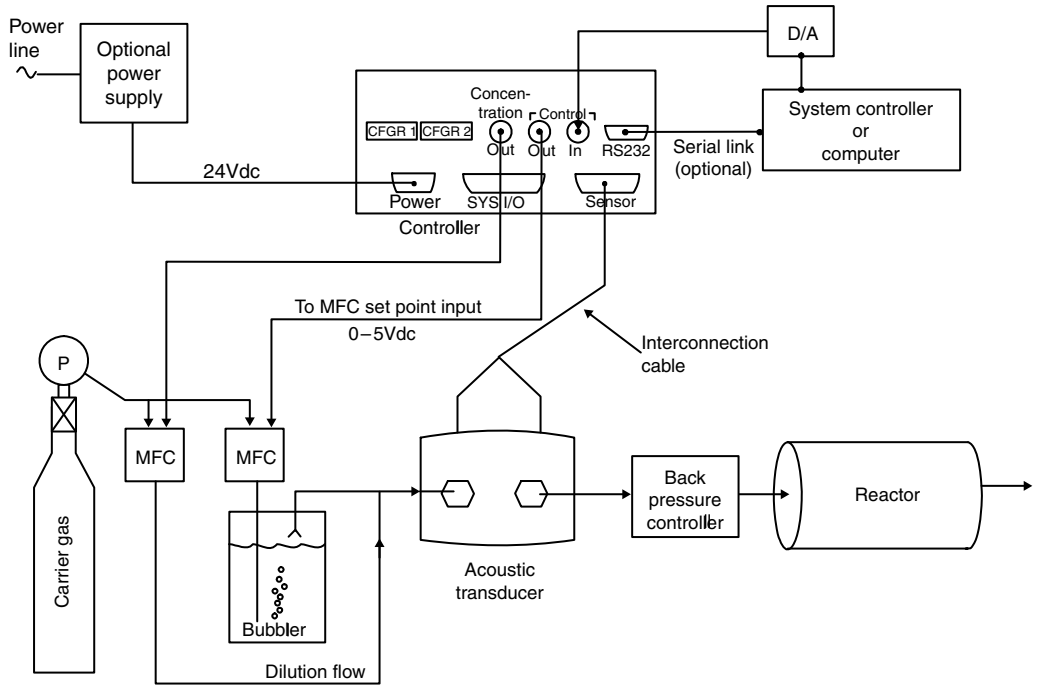


FIGURE 25.13 Typical installation of an acoustic composition measurement system. (Adapted from input by Gogol, C.A., Leybold Inficon, East Syracuse, NY. http://www.inficon.com)

There is little effect on the gas supply dynamics as either implementation adds little additional volume ($<25\ cm^3$) to the reactor's delivery system.

A typical installation of an acoustic composition measuring and control system [41] is shown in Figure 25.13. It consists of two major components. First, a transducer is directly inserted into the reactor's supply line. Second, an electronic control console is used for controlling the sensor's temperature, determining the speed of sound, computing the composition, generating feedback control voltages for mass flow sensors, and analog and digital transmission of relevant process control data. The gas delivery tube is cut and reconnected so the gas mixture passes through the transducer. The transducer's temperature may be set to match the transport needs of the materials or match heated delivery tubes. This installation demonstrates simultaneous control of both the bubbler flow and the dilution flow, thus maintaining constant composition at constant total flow.

25.2.2.4.3 Sensor Sensitivity, Stability, and Calibration

Figure 25.14 demonstrates how the speed of sound varies with relative composition for some common gas pairs. The steep slope at low concentrations of a high molecular weight gas in a light carrier allows concentrations as small as 1 ppm to be measured. The sensitivity of these techniques is strongly influenced by the difference in mass between the species and the particular range of compositions. The technique is most sensitive for low concentrations (less than 5 mol%) of a high molecular weight species in a light gas. Even when the molecular weight differences are small, e.g., O_3 in O_2 or N_2 in Ar, it is generally easy to discern compositions differing by 0.1% or less for all concentrations.

Acoustic analysis is stable and highly reproducible over long periods of time. Reproducibility can be further improved with daily calibration. Calibration is simple if the installation of the sensor permits pure carrier gas to flow through the sensor. Calibration is the renormalization of the instrument's

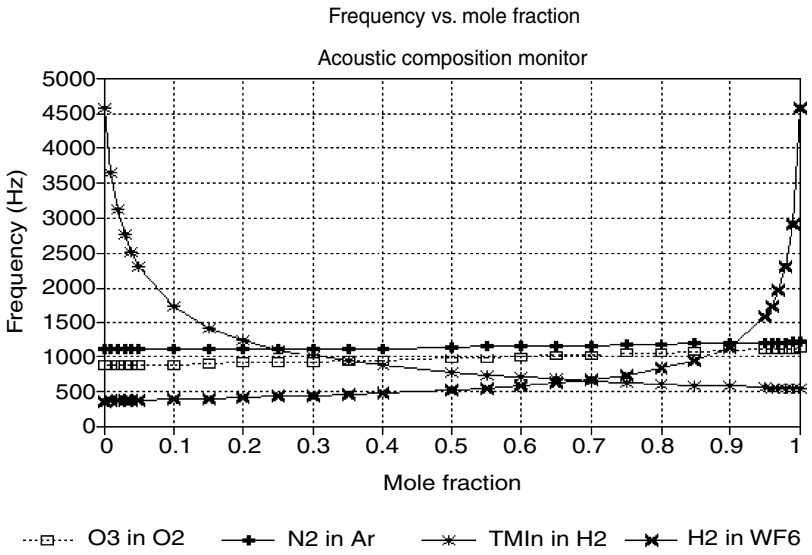


FIGURE 25.14 Graph of frequency vs. mole fraction for various binary gas mixtures. (Adapted from input by C.A. Gogol, Leybold Inficon, East Syracuse, NY. http://www.inficon.com)

effective path length, *L*. This is easily accomplished at the point of installation by measuring a known pure gas, generally the carrier gas. The calibration process is brief, only requiring the sensor and its supply lines to be sufficiently flushed in order to dilute any remaining precursor so that it causes no measurable effect on the speed of sound. This condition is readily observable because the speed of sound will asymptotically stabilize once the other gasses are sufficiently flushed.

Lower temperatures and higher pressures are more easily measured with acoustic techniques as acoustic transmission is dependent on media density. At low gas density, it is difficult to transmit sufficient sound intensity to overcome the parasitic sounds transmitted through the sensor structure itself or to distinguish the signal from the other sounds in the environment. These problems have presently limited the successful operation of this technology to situations where the supply gas already exceeds or can be compressed to pressures over 70 Torr.

25.2.2.4.4 Sensor Integration

Integration of this sensor with the reactor may be either analog or digitally based. These sensors' electronics implementation is dominantly digital, but the composition signal is also available on a precision analog output which allows the concentration to be interpreted by precise analog readout. The preferred interfacing is digital, as the transducer's inherent precision and dynamic range exceed that normally carried as an analog process control signal. A digital interface allows a graphic display to be used to advantage to uncover process flaws, such as improper reactor gas line switching and to readily view the concentration instabilities that indicate source depletion. It is also straightforward to save a detailed record of each process run when utilizing a digital interface.

The future direction for the development of this instrumentation is to enhance operation from the present operating limit of 65°C and allow its use at temperatures in excess of 120°C. Another productive path will be to learn how to apply these simple and robust transducers in networked arrangements. It is felt that they will have measurement capabilities beyond the present 2-gas limit and may reliably provide information on reactor efficiency and perhaps infer wafer properties such as the rate of film growth.

25.2.3 RF Properties

A significant number of etch and deposition tools are RF powered. For these tools, the process state includes the characteristics of the RF conditions, such as the delivered power or the current, voltage, and phase information—at the fundamental frequency and the higher harmonics. The major benefit of RF sensors is that they are readily available, relatively inexpensive, and readily couple into the RF circuit. Their major drawback is that the accuracy of the measured parameters is a complex function of the placement and impedance matching of the sensor in the RF path. Specifically, any RF sensor of fixed impedance, placed between the matching network and the dynamically changing impedance of the plasma chamber will generate measurements of questionable accuracy. Calibrating the sensor for variable load impedances can mitigate this issue. But viewed pragmatically, this complexity is the reason why commercial plasma tools are not yet controlled by postmatch RF sensors. If accuracy is not the primary requirement, these sensors can generate useful data for endpoint determination or fault detection.

25.2.3.1 Sensor Technologies

An RF sensor [42] is a device that produces output signal(s) that are of a definite and defined relationship to the electrical energy present in or passing through the sensor. To allow for the placement of sensing elements into controlled and reproducible electromagnetic field conditions, RF sensors are typically designed and built around transmission line structures.

Minimum system disruption by the RF sensor is important to guarantee that the load seen by the source network is the same with or without the RF sensor. In other words, the measurement device should not significantly change the load it is trying to measure. A typical RF sensor will produce the following insertion disruptions:

1. Capacitance to ground,
2. Series inductance.

A small capacitance to ground is essential in preventing the sensor from increasing the reactance of the load as seen by the source network. The series inductance is generally designed in combination with the capacitance to ground to produce the characteristic operating impedance of the sensor (usually 50 Ω for RF applications). A small series resistance allows the sensor to have low insertion loss (i.e., dissipating power in the sensor instead of the load). Having a small value for the series resistance is crucial to maintaining a high Q (quality factor) of the load network and allowing for high system efficiencies. The following describe the two major types of RF sensors: directional couplers and voltage/current sensors.

Directional coupler is the most common type of RF sensor. It is generally used to measure the forward and reverse power at the generator output by directionally sensing the RF components at the sensor. These values are generally accurate, as the 50 Ω sensor is connected to the stable 50 Ω input to the matching network.

Commercially available directional couplers are typically rated in terms of the forward and reverse power, the sensing element can withstand and usually have a specific coupling coefficient (e.g., −30 dB) over a specified frequency bandwidth and a characteristic impedance (typically 50 Ω). Directional couplers are available from a number of vendors [43,44].

VI sensors are the second most common types of RF sensor. This sensor's operation relies upon the electrical principles of capacitance and mutual inductance. A capacitor is formed when two conductors are placed in parallel to one another and separated by a certain distance. If the output of the capacitor is connected to a suitable shaping network, the voltage drop across the created capacitor can be controlled to produce an output signal with repeatable attenuation. The ideal capacitive structure, when built into a transmission line, is as wide and short as possible. This allows for maximum voltage coupling (by maximizing the capacitance) and minimum current coupling (by minimizing the mutual inductance) into the sensor device network. A mutual inductor is easily formed when two conductors are placed in parallel with each other. An ac current traveling in one conductor will produce an ac current in the other conductor traveling 180° out of phase with it. The ideal inductive structure, when built into a

transmission line, is as long and thin as possible. This allows for maximum current coupling (by maximizing the mutual inductance) and minimum voltage coupling (by minimizing the capacitance) into the sensor device network. It is important to be aware of a possible current contamination in the voltage sensor and voltage contamination in the current sensor. If this occurs, the dynamic impedance range and maximum sensor accuracy will be sacrificed.

One important thing to recognize when using voltage and current sensors is that each independent sensor must look at the same position on the coaxial line. Also consider that the sensor and the associated cables and electronics have been specifically calibrated as a unit; hence, no component can be arbitrarily changed without recalibration. If these rules are not followed, the standing wave ratio seen by each sensor will be different—allowing for errors in the produced signals.

Voltage and current sensors are typically rated in terms of how many amperes or volts the sensor can tolerate. This rating is similar to the transformer rating specified in terms of VA (volts$\times$amperes). The VI sensors are also commercially available [45–48].

25.2.3.2 Measurement Technologies

A measurement technology is necessary to process the outputs of the RF sensor. In the past, measurement techniques have been typically analog-based signal processing. Since the advent of the DSP, more and more measurement techniques have migrated to the digital world. For any type of measurement technique to perform well, it must have the following minimum characteristics:

- reproducible results—stable vs. time and environmental conditions;
- wide frequency range;
- wide sensitivity range;
- impedance independent accuracy;
- $\pm 180^\circ$ phase measurement capability;
- flexible calibration and calculation algorithms.

Having a measurement technique with reproducible results is a must for any sensor system. Day-to-day reproducibility allows for maximum reliability of the sensor, while unit-to-unit reproducibility allows for data interpretation to be consistent for each unit purchased. An excellent unit-to-unit reproducibility is absolutely necessary if a sensor system is to be used in a manufacturing environment. Inherent in reproducibility is low drift. Low drift overtime in a sensor system's readings is necessary for day-to-day and measurement-to-measurement reproducibility. Also, because of the large temperature ranges produced by many of the new plasma processes, low temperature drift is necessary to maintain maximum accuracy.

Many single frequency sensor systems are available on the market today, but a sensor system with a measurement technology that performs over a wide frequency range allows the user to look at harmonics (for a single frequency processes) and mixing products (for multiple frequency processes) without incurring additional cost. Hence, a sensor system with a wide frequency range has the lowest cost of ownership.

Especially, if the sensor is used over a wide frequency range, a wide range of sensitivity is required. The magnitudes of the signals at the fundamental vs. the upper harmonics can be significantly different, hence requiring a large dynamic range in the sensor sensitivity.

Some sensor systems have accuracy specifications that depend upon the impedance of the load. For maximum reproducible accuracy, a sensor system that uses a measurement technology with impedance independent accuracy must be employed. The most important values to be measured are the fundamental electrical parameters of $|V|$, $|I|$, and $< Z$ (the phase angle of the load, or the phase angle between the voltage and the current). These three parameters are the building blocks of all other electrical parameters (such as power, impedance, reflection coefficient, etc.). Some sensor system vendors specify their accuracy in terms of nonelemental parameters; in this case, a little algebra is necessary to transform the specifications to the elemental parameters.

Passive loads (formed with capacitors, inductors, and resistors) can only result in impedance phase angles in the $\pm 90^\circ$ range, while active loads can produce any phase angle over the $\pm 180^\circ$ range. Due to the complicated physical processes that govern electron and ion transport in a plasma, the resulting electrical impedance produced by the plasma is active. Hence, to allow for proper measurement of a plasma load, the sensor system must be capable of determining phase angles in the $\pm 180^\circ$ range—all possible phase angles.

Another consideration for a sensor system is its upgrade path. Typical analog techniques process, the sensor signals with circuitry. Due to the fact that any technology improvement requires circuit redesign, analog processing does not allow for a low cost upgrade path for technology improvements. Hence, the lowest cost of ownership in a sensor design is achieved with a digital technique that allows for signal processing upgrades with new versions of software.

25.2.3.3 Signal Processing

Once the sensor signal is obtained (see Section 25.2.3.2), it has to be processed to derive the parameters of interest. In some cases, signal processing requires the down-conversion of the RF signals to a lower frequency that is more easily digitized. Once in the digital domain, DSP algorithms provide a very efficient and flexible way to process these sensor signals. In contrast to available analog signal processing methods, digital signal processing is done completely with software, not hardware. Hence, the flexibility of calculation and calibration algorithms is very high. Any improvements to sensor or calculation technology can be implemented in software—drastically reducing the design cycle for improvements in the signal processing technology. Another important advantage of having a DSP-based embedded system in the design is completely self-contained operation. Additional hardware is not necessary to support operation of the unit because all calibration information can be stored in DSP nonvolatile memory. In addition, the DSP can allow for user selectable high speed filtering of data.

A RF sensor system should be able to extract the following data at the frequency of interest:

$\|V\|$	Root mean square (RMS) voltage	V
$\|I\|$	RMS current	A
$\|Z\|$	Impedance magnitude of load	W
θ	Phase angle of load	Degrees or radians
P_D	Delivered (load) power	W
P_F	Forward power	W
P_R	Reverse power	W
P_{RE}	Reactive (imaginary) power	VAR
Γ	Reflection coefficient	No unit

Due to the mathematical relationships of the above nine parameters to each other, the RF sensor system must be able to directly measure three out of the nine parameters to properly calculate the remaining six. The accuracy with which each of these three fundamental parameters is measured determines the accuracy to which the other six parameters can be calculated and overall sensor system quality. A broadband RF sensor system will allow the user to extract data at harmonics to more thoroughly characterize the behavior of the RF plasma and RF system.

25.2.3.4 Sensor Installation and Use

The two typical installations of a RF sensor system are shown in Figure 25.15 and Figure 25.16.

As demonstrated, the RF sensor can be mounted either before or after the matching network. One thing to realize is that any such 50 Ω sensor will perturb the $V|I|$phase values that existed in a non-50 Ω path without the sensor in place. Impedance mismatch between the sensor and the point where it is inserted will generate RF reflections; thereby influencing the RF environment. This does not negate their utility, but one needs to consider that the measurement itself changes the RF environment. The second issue is that, whether the sensor is located pre- or post-match, it reads the instantaneous $V|I|$ phase values at that point along the transmission path. These values are indicative of the standing wave characteristics at that point

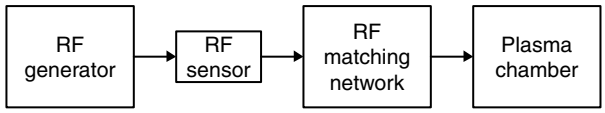


FIGURE 25.15 RF sensor mounting, pre-match. (Adapted from input by Kevin S., Gerrish, ENI Technology, Inc., Rochester, NY.)

in the transmission path. However, these values will be influenced by the plasma properties, which is the primary reason for the use of these sensors for endpoint or fault detection. The changing impedance of the plasma creates changes in the standing wave characteristics along the transmission path, most dramatically between the tuner and the plasma. Hence these sensors, located either pre- or post-match, will see changes in the plasma. One benefit for locating sensors pre-match is the relative ease of mounting the sensor with standard coaxial coupling, assuming that a useful signal can be obtained in this location.

The analysis and interpretation of the available sensor data require that one comprehend that the sensor measures the instantaneous $V|I|$ phase values at one specific location in the RF transmission path. What happens at another location (namely the plasma) can be inferred by correlation (i.e., a change in the standard measured values) or by means of a full RF-circuit model. Such models are generally very difficult to generate; hence, the majority of the RF sensor data analysis is performed by the simpler correlative method.

25.2.3.5 Applications of an RF Sensor System

In spite of the previously described limitations, RF sensors can be gainfully utilized in a number of applications. In some cases, they are relatively "easy" and inexpensive add-on sensors, and have shown benefits in applications where accuracy is not a key parameter (as long as sensor reproducibility persists). The following sections describe the examples of these applications.

25.2.3.5.1 Etching Endpoint Detection, Deposition Thickness

For this application, the RF sensor system can be mounted before or after the RF matching network. Even in a pre-match location, an RF sensor system with enough sensitivity can detect the small variation in plasma impedance that depicts an etching endpoint or accompanies a deposition process. Using VI sensors in a pre-match configuration on a plasma etcher, endpoint signals have been seen in the higher harmonics [85]. For an oxide deposition, a capacitance change will also be seen by the RF sensor. The value of the observed capacitance could be correlated to the thickness of the deposited film. Similar information can be obtained from sensors placed post-match.

25.2.3.5.2 Harmonic Signature Analysis

Due to the nonlinear characteristics of the plasma, the pure sine wave from the RF generator will be turned into a harmonic-rich waveform by the plasma. The number of RF harmonics present, as well as the characteristics of each, will depend on the plasma chamber geometry, the type of energy source for the plasma chamber, and the type of process being run.

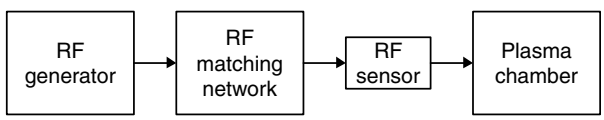


FIGURE 25.16 RF sensor mounting, post-match. (Adapted from input by Kevin S., Gerrish, ENI Technology, Inc., Rochester, NY.)

Proper use of this technique would create a "harmonic fingerprint" of the process when it is running well. Future process fingerprints would be compared, probably by a multivariate numerical technique, to the master fingerprint at regular intervals. Any significant change between the two would indicate a process shift, allowing the chamber to be taken off-line before a complete lot of wafers is destroyed. If enough data are taken, a database could be created allowing proper interpretation of the bad harmonic fingerprint. Examples of anomalies expected to be found by this technique are broken wafer, misplaced wafer, dirty chamber (i.e., chamber cleaning required).

It is wise not to limit the harmonic range. Experiments have indicated [49] that higher harmonics (10th through the 20th) can contain stronger process correlation than lower harmonics, as well as different information. A particular RF investigation on a cluster-tool polysilicon etcher chamber B, it was found that the seventh harmonic, or 94.92 MHz, had a good etch endpoint trace. Looking at the higher harmonics, the 13th harmonic, or 176.28 MHz, showed a strong reaction to chamber D's RF power cycle. This indicated that the 13th harmonic should not be used to characterize chamber B. At the 16th harmonic, or 216.96 MHz, an endpoint signal was found with much better endpoint (EPT) characteristics than the seventh harmonic. No other harmonics, up to the 20th, produced any usable EPT information.

25.2.3.5.3 Measurement of Power Delivered to Plasma

The more difficult application is the accurate measurement and control of power delivered to the plasma. A typical RF system will regulate the output power of the RF generator to very high accuracy. Unfortunately, every RF delivery system has losses. In most cases, the losses change as a function of generator power due to plasma impedance changes. Also, the losses in a RF delivery system may increase as the system ages (e.g., wear of the mechanical components in the tuner). This means that the actual power delivered to the plasma is always less than the output power of the generator, and may change from wafer-to-wafer and lot-to-lot. *A properly designed and calibrated* RF sensor connected between the matching network and the plasma chamber allows for measurement of power delivered to the plasma. With valid measurement of the true delivered RF power, the RF sensor system can be used to compensate for the losses described above and decrease wafer-to-wafer processing variations. This approach could provide a better tuning algorithm using the impedance information from the RF sensor to correctly set the capacitor values of the RF matching network; although additional feedback will be required to prevent excessive reflected power from damaging the generator. This is all leading towards the implementation of a feedback loop from the post-match sensor to the RF generator. The goal is to provide more consistent and accurate RF power delivered to the plasma in commercial plasma tools.

25.2.4 Wall Deposition Sensor

Process state sensors have classically focused on determining the species concentrations, pressure, gas flow, and RF power characteristics in the processing environment during the processing of each wafer. In the recent quest for more control over the reproducibility of processing tools, attention has recently been focused on the deposits generated on the internal surfaces of processing tools. Such deposits are formed in both deposition and etch tools; typically at a greater rate in deposition systems. These deposits have several detrimental effects:

- Provide a slowly changing chamber-wall state, which can influence the process chemistry.
- At some point, the deposit can start to flake off the walls and potentially contaminate the wafer.

The last point drives the mechanical and plasma cleaning of tools, in hopes of preventing this source of particle contamination. Without an appropriate sensor, the frequency of these cleaning cycles is based on empirical rules; with an incorrect guess risking either wafer contamination or expensive, unnecessary downtime for tool cleaning.

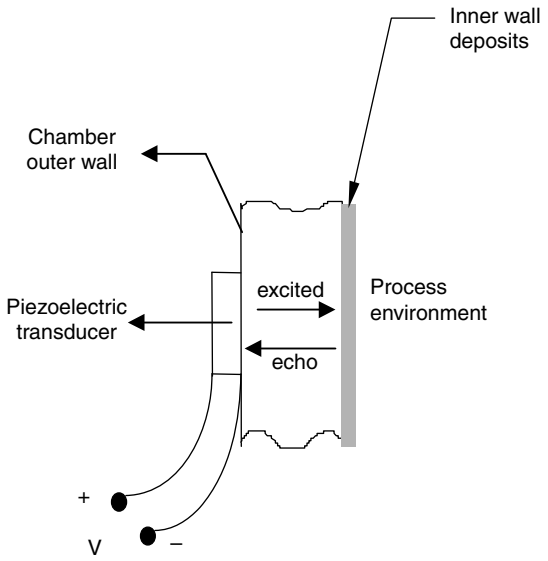


FIGURE 25.17 Cross-sectional view of a chamber with the piezoelectric transducer attached to the outer wall.

25.2.4.1 Theory of Operation

A sensor has recently been developed [17] for real time, noninvasive monitoring of chamber wall deposits in etch and deposition process tools. This sensor [50] can be used in an R&D or production environment for optimizing clean cycles and reducing particle contamination. It operates on the principle of acoustic reflectometry (Figure 25.17). A piezoelectric transducer is attached to the outer wall surface of a process chamber. A short electric pulse applied to the electrodes of the piezoelectric transducer excites an acoustic wave that propagates from the outer wall of the chamber toward the inner wall. If the inner wall is bare (with no deposits), the acoustic wave is reflected as an echo from the boundary between the inner wall and the process environment. This echo propagates to the outer wall where the transducer converts it into a detectable electrical signal. The transit time of the round trip by the acoustic wave is the fundamental measurement. When a film is deposited on the inner wall, any change in the thickness of the deposit causes a proportional change in the transit time. A change in temperature also changes the transit time, but in a very predictable manner. Hence the sensor acoustically monitors, in real time, changes in the average temperature along the cross-section of the chamber wall, with a real-time temperature compensation applied to the measurement described above.

The sensor system consists of a personal computer that houses the electronics and the transducer module. The transducer module is attached to the outer wall (or window) at the chosen location on the process chamber. The transducer module is a cylinder 2 in. in diameter and 2 in. in height. The primary use of this sensor is for determining and optimizing the chemistry, frequency, and duration of clean cycles for etch and deposition tools.

25.3 Wafer-State Sensors

As stated previously, process state sensors have predominantly been used for endpoint determination and fault detection, and in some recent cases for dynamic process control. But clearly, wafer-state sensors provide more direct information for all these tasks. Such wafer-state sensors are slowly being integrated into processing tools, paced by issues of: customer pull, sensor reliability, cost of integration, etc.

The following is a description of the wafer-state sensors that have, or are currently overcoming these barriers and are being integrated into OEM tools.

25.3.1 Film Thickness and Uniformity

The thickness of optically transparent thin films (silicon, dielectrics, resists) on a reflective substrate is measured via the analysis of the interaction of electromagnetic radiation with such a film, or film stack. These methods rely on single wavelength (laser) or spectral (white light) sources, impinging on the sample at normal incidence (interferometry and reflectometry) or at some angle off-normal (reflectometry, ellipsometry). The wavelength range is from the UV through the IR. The interaction of the light with the material can be detected through a polarization change (ellipsometry), a change in the phase (interferometry), or a change in the reflected amplitude (reflectometry). Optical models are used to extract the different physical parameters of the films (e.g., thickness) from the known optical indices of the individual layers. These techniques are well-established standard methods for off-line film thickness measurement, and hence the methods will only be briefly described. The emphasis will be on the deployment of these techniques as sensors in OEM tools.

25.3.1.1 Optical Sensors

The spectral reflectivity of transparent thin films on reflective substrate materials is modulated by optical interference. The effect of the interference on the measured spectrum is a function of the film and the substrate refractive indices. If the dispersion components of the refractive indices are known over the wavelength range, the thickness of the surface film can be found using a Fourier transform technique. For thin layers (<100 nm), the method of spectral fitting is very effective. Once the film thickness has been found, a theoretical reflectance spectrum can be determined and superimposed on the measured spectrum. This ensures a very high level of reliability for the film thickness measurement.

25.3.1.1.1 Reflectometry Technique

25.3.1.1.1.1 Theory of Operation

The thickness of films on a silicon wafer is measured by means of spectrophotometry, utilizing the theory of interference in thin films [51]. The basic procedure is to measure the spectral reflectance of the desired sample. The spectral data are then interpreted to determine the thickness of the top layer of the measured stack. The actual reflectance $R_{act}(\lambda)$ is measured and fitted to $R_{theor}(\lambda)$ to find the thickness (d) of the last layer. $R_{theor}(\lambda)$ is calculated according to the specific optical model of the measured stack. The "goodness-of-fit" parameter measures the difference between the measured and the theoretical results and is used as a criterion of correct interpretation. Figure 25.18 shows a graph of $R_{theor}(\lambda)$ for a layer of 10,000 Å SiO_2 on Si substrate.

The fitting algorithm used for data processing has to treat several issues such as spectral calibration, noise filtering, recognition of characterizing points (minima, maxima, etc.) and calculating a first-order approximation for the thickness and the final fine fitting.

25.3.1.1.1.2 Optical Overview

The optical path of one specific reflectometer [52] is shown in Figure 25.19. In this case, the specular reflection is monitored at the incident angle normal to the wafer surface, and the radiation source is in the visible range.

Briefly, the light emitted from the lamp (11) travels through an optical fiber (10) until reaching a condenser lens (9). The light beam then reaches a beam splitter (3) where it is split; half of the light passes through the beam splitter, while the other half is reflected downwards, focused by a tube lens (2) and an objective lens (1) onto the target (wafer). After being reflected by the target, the light beam travels back

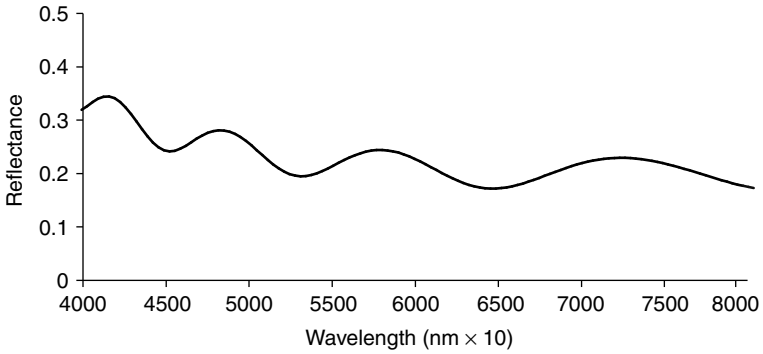


FIGURE 25.18 Reflectance of SiO_2 on Si in water. (Adapted from input by Ran Kipper, Nova Measuring Instruments Ltd., Weizman Scientific Park, Rehovoth, Israel.)

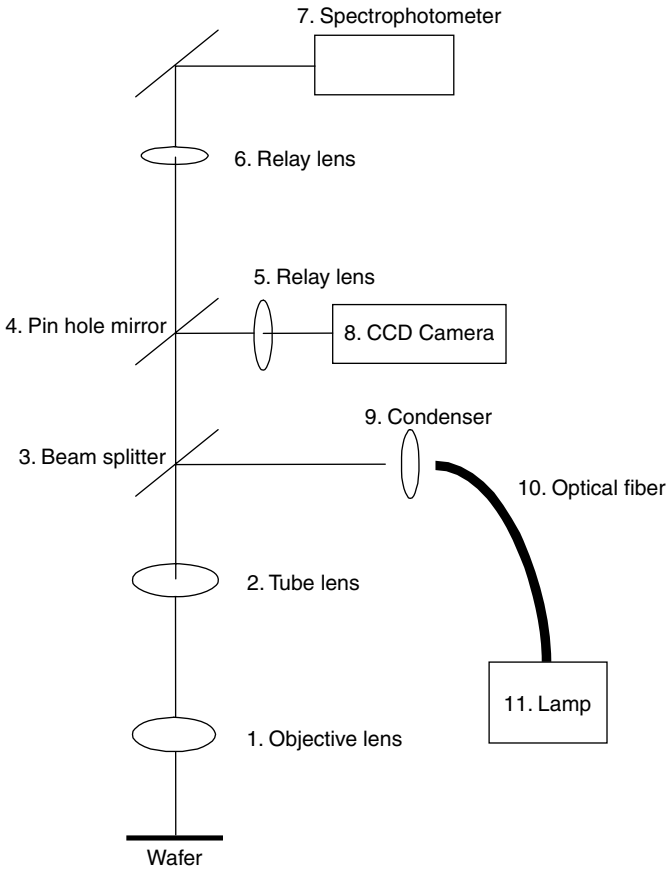


FIGURE 25.19 Optical path of light beam in NovaScan 210. (Adapted from input by Ran Kipper, Nova Measuring Instruments Ltd., Weizman Scientific Park, Rehovoth, Israel.)

through the objective lens (1), tube lens (2), and through the beam splitter (3) until it reaches a "pin-hole" mirror (4). From there, the light is sent in two directions:

a. A portion of the light (the image of the wafer surface) is reflected from the "pin-hole" mirror (4), focused by a relay lens (5) onto a CCD camera (8) where it is processed and sent to the monitor for viewing by the operator.
b. The light that passes through the "pin-hole" is also focused by a relay lens (6), then reflected by a flat mirror towards the spectrophotometer (7), which measures the spectrum of the desired point. This information is then digitized and processed by the computer for the computation of film thickness.

The above spectrophotometer also includes an auto-focusing sensor for dynamic focusing on the wafer surface during the movement of the optical head over the wafer.

25.3.1.1.1.3 System Integration, In-Line Measurement

While this chapter is focused on in-situ metrology, there are some well-established in-line measurement techniques that are worth including as they provide routine and useful information for APC (specifically, wafer-to-wafer control). Two embodiments of in-line reflectometry for film thickness measurements will be described in this section; one for use in chemical–mechanical polishing (CMP) and the other for epi film growth.

Reflectometry is used to monitor and control film thickness in CMP operations. When CMP is used to planarize and remove part of a blanket film, such as in oxide CMP, there is no detectable endpoint since no new films are exposed. The only way to monitor and control such a process is by a sensor that measures the thickness of the film. This is a very difficult task for a slurry-covered wafer that is in motion, hence the measurement is performed in-line in the rinse station of the CMP tool.

A commercially available reflectometry-based sensor [52] is currently being used for CMP tool monitoring. Its primary benefits are as follows:

- Provides thickness measurement data for every product wafer, required for rapid feedback control of the CMP process.
- Performs measurements in parallel to the processing of the next wafer, hence not affecting system throughput unless a very large number of measurements are required.
- In-water measurement capability obviates the need to clean and dry wafers before measurements.
- Additional clean-room space and labor required for off-line measurements are eliminated.

Only one component, the measurement unit, has to be integrated into the polisher. The compact size of this unit, with a footprint only ~40% larger than the wafer, enables easy integration into the process equipment.

Two such implementations in commercial CMP tools are represented in Figure 25.20 and Figure 25.21.

Two different delivery system principles are applied for the integration of the measurement system into OEM tools. In one case (Figure 25.20), the wafer handler transfers wafers down from the wafer loading station to the water tub of the measuring unit and back. In another configuration (Figure 25.21), the measurement unit replaces the unload water track of the polisher. It receives the wafer, performs the measurement process, and delivers the wafer to the unload cassette. In both cases, the wafer is wet during the measurement.

A second commercially available implementation of reflectometry (in this case using an IR source and non-normal incidence) is the use of FTIR measurement of epi thickness. The in-line measurement of epi thickness has been achieved by the integration of a compact FTIR spectrometer [53] to an Applied Materials Epi Centura Cluster tool, as shown in Figure 25.22. The cool down chamber top plate is modified to install a CaF_2 IR transparent window, and the FTIR and transfer optics are bolted to the top plate. The IR beam from the FTIR is focused to a 5-mm spot on the wafer surface, and the specular reflection is collected and focused onto a thermoelectrically cooled MCT detector. Reflectance spectra can be collected in less than 1 s. Reference spectra are obtained using a bare silicon wafer surface mounted within the cool-down chamber. Epi thickness measurements are made after processing, while the wafers

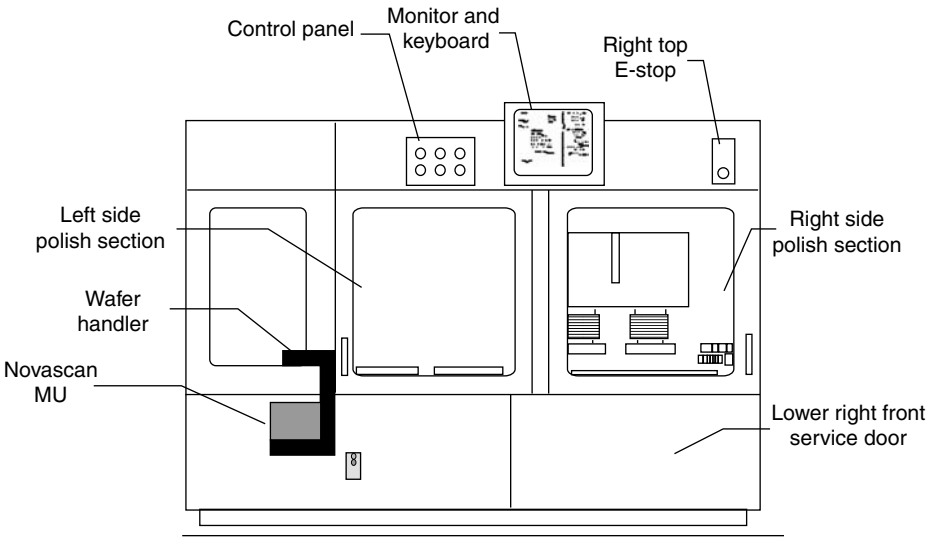


FIGURE 25.20 NovaScan system integrated in Strasbaugh Model 6DS-SP Planarizer. (Adapted from input by Ran Kipper, Nova Measuring Instruments Ltd., Weizman Scientific Park, Rehovoth, Israel.)

are temporarily parked in the Cluster Tool's cool-down chamber, without interrupting or delaying the wafer flow.

A simulated reflectance spectrum is computed from parametric models for the doping profile, the dielectric functions (DFs) of the epi film and the substrate, and a multilayer reflectance model. The models for the wavelength-dependent complex DFs include dispersion and absorption due to free-carriers, phonons, impurities, and interband transitions. The models are tailored to the unique optical and electronic properties of each material. The reflectance model computes the infrared reflectance of films with multilayered and graded compositional profiles using a transfer matrix formalism [53,54]. The model parameters are iteratively adjusted to fit the measured spectrum.

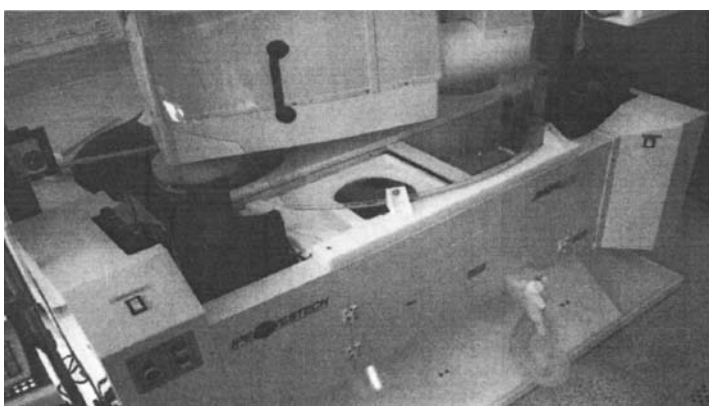


FIGURE 25.21 NovaScan in IPEC 372M and 472 Polisher. (Adapted from input by Ran Kipper, Nova Measuring Instruments Ltd., Weizman Scientific Park, Rehovoth, Israel.)

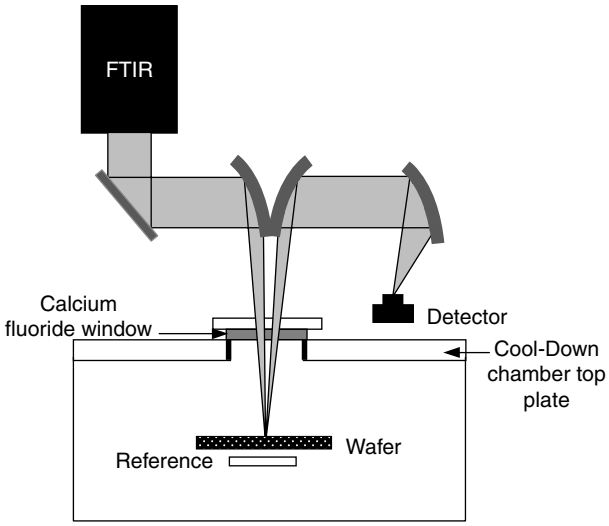


FIGURE 25.22 Configuration of On-Line Technologies, Inc. FTIR on Applied Materials' Centura 5200. (From NovaScan, Nova Measuring Instruments Ltd., Weizman Scientific Park, Rehovoth, Israel.)

Gauge tests demonstrate the relative accuracy of this first principle analysis of epi layer thickness to be in the range of 0.5–2 nm (5–20 Å). Comparison to destructive secondary ion mass spectroscopy (SIMS) and spreading resistance analysis (SRP) measurements shows the absolute accuracy to be within the accuracy of these standard measurements.

25.3.1.1.2 Interferometric Technique

25.3.1.1.2.1 Theory of Operation

Interferometry is a well-established technique for the optical measurement of thin, optically transparent films. Some of the light impinging on such a thin film reflects from the top of the film and some from the bottom of the film. The light reflected from the bottom travels farther and the difference in path length results in a difference in phase. After reflection, the light following the two paths recombines and interferes, with the resulting light intensity a periodic function of the film thickness. The change in film thickness for one interferometric cycle is $\lambda/2n \cos\theta$, where λ is the observation wavelength, n is the index of refraction of the film, and θ is the angle of refraction within the film.

25.3.1.1.2.2 Full Wafer Imaging Sensor [55]

The full wafer imaging (FWI) sensor is a novel sensor developed in the early 1990s [57] based on this interferometric technique. It uses an imaging detector to make spatially resolved measurements of the light reflected from the wafer surface during etching or deposition processes. This sensor takes advantage of the fact that the reflectivity of a thin film on the wafer surface is generally a function of the thickness of the film. By quantifying the changes in reflectivity as the film thickness changes, the FWI sensor determines spatially resolved etching or deposition rate, rate uniformity, spatially resolved endpoint, endpoint uniformity, and selectivity.

These measurements are performed on every wafer, providing both real-time endpoint and run-by-run data for process monitoring and control.

The operation of this particular sensor relies on a number of optical phenomena:

- *Optical emission*. Optical emission from the plasma is the preferred light source for FWI sensors, because it is simpler than using an external light source and it allows direct detection of optical emission endpoint. If plasma light is not available, an external light source can be added.

A narrow bandpass filter is used to select the measurement wavelength. Different wavelengths are best suited to different types of process conditions; the most important characteristics being the intensity of the plasma optical emission as a function of the wavelength, the film thickness, and the film's index of refraction. In general, a shorter wavelength gives better rate resolution, but cannot be used in certain situations, e.g., a 0.3-μm thick layer of amorphous silicon is typically opaque in the blue, but transparent in the red.

- *Interferometry for transparent thin films*. In practice, during an etching or deposition process, the intensity of light reflected from the wafer surface varies periodically in time. The interferometric signal is nearly periodic in time in most processes because the process rate is nearly constant, even though the signal is strictly periodic in film thickness rather than in time.
- *Interferometry for trench etching*. In trench etching, the interference is between light reflected from the top of the substrate or mask and light reflected from the bottom of the trench. A coherent light source, e.g., a laser, must be used because the interference is between two spatially distinct positions. Etching rate is calculated using the same types of techniques discussed above for thin film interferometry. Endpoint time is predicted by dividing the desired trench depth by the measured etching rate.
- *Reflectometry for nontransparent films*. Light impinging on a nontransparent film reflects only from the top of the film, so there is no interference. However, the reflectivity of the nontransparent film that is being etched is different from the reflectivity of the underlying material. Thus, the intensity of reflected light changes at endpoint. This method is typically applied to endpoint detection in metal etching.

From a system viewpoint, the FWI sensor requires a high data acquisition rate and uses computationally intensive analyses. So the typical configuration consists of a high end PC, advanced software, and one or more independent CCD-based sensor heads interfaced to the computer via the peripheral component interconnect (PCI) bus. Each sensor head records images of a wafer during processing, with each of the few hundred thousand pixels of the CCD acting as an independent detector. The full images provide visual information about the wafer and the process, while the signals from thousands of detectors provide quantitative determination of endpoint, etching or deposition rate, and uniformity. The simultaneous use of thousands of independent detectors greatly enhances accuracy and reliability through the use of statistical methods. The FWI sensor can be connected to sensor bus by adding a card to the PC. Connecting the sensor head directly to sensor bus is not practical, due to the high data rate and large amount of computation.

Figure 25.23 shows a schematic diagram of the FWI sensor head installation. The sensor head is mounted directly onto a semiconductor etching or deposition tool on a window that provides a view of the wafer during processing. A top-down view is not necessary, but mounting the sensor nearly parallel to the wafer surface is undesirable, because it greatly reduces spatial resolution, one of the technique's principle benefits.

For both interferometry and reflectometry, spatially resolved results are determined by applying the same calculation method to hundreds or thousands of locations distributed across the wafer surface. These results are used to generate full wafer maps and/or to calculate statistics for the entire wafer, such as average and uniformity. Several methods can be used to find the etching or deposition rate from the periodic interference signal.

The simplest way is to count peaks, but this is accurate only if there are a large number of interferometric cycles, which is not common in most semiconductor processes. For example, a 0.3-μm thick layer of polysilicon contains only 3.8 interferometric cycles. The accuracy of simple peak counting is one-half of a cycle, which is only 13% in this example. The accuracy can be improved somewhat by interpolating between peaks, but the accuracy is still fairly low. In addition, false peaks caused by noise in the signal often plague peak counting methods.

A more accurate way to determine rate is to multiply the change in film thickness per interferometric cycle by the frequency (number of cycles per second). The simplest way to find the desired frequency is to

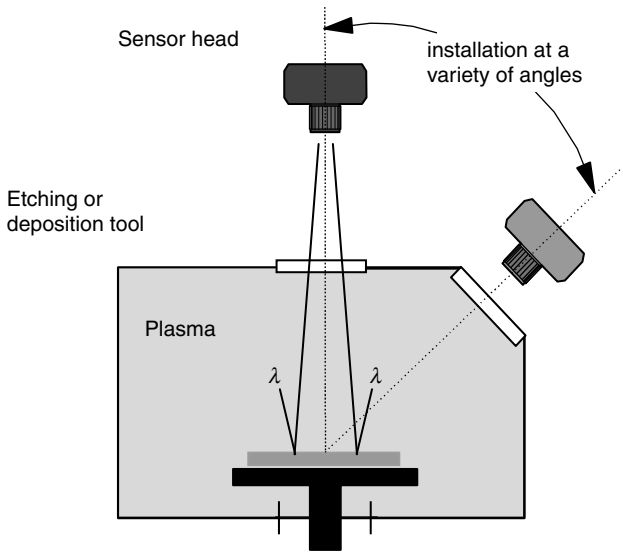


FIGURE 25.23 Full wafer imaging (FWI) sensor head mounting. (Adapted from input by Conner, W.T., Leybold Inficon, Inc., East Syracuse, NY. http://www.inficon.com)

use a fast Fourier transform (FFT) to convert from the time domain to the frequency domain. A local maximum in the signal vs. frequency then specifies the frequency to be used in the rate calculation. Accuracy can be further increased by starting with an FFT to provide an initial guess of the frequency and then fitting the signal vs. time to an empirical function that models the physical signal. This combined method is more accurate than the FFT alone if there are few cycles, if the interferometric signal is not a pure sine wave, or if the SNR is low—all of which occur commonly in semiconductor processing. In either method, a frequency window can be used to designate, which maximum in the FFT is used to calculate rate. This is a useful way to measure selectivity in etching processes where two materials, e.g., the mask and the film of interest, are etching simultaneously.

For transparent thin films, endpoint can be detected or predicted. The detection method relies on the fact that the periodic modulation of reflected light intensity ceases at endpoint. Endpoint is found by detecting the deviation of the observed signal from an interferometric model. The prediction method uses the measured rate and the desired thickness change to predict the endpoint time. Prediction is the only available endpoint method for deposition processes. It is also necessary in those etching processes where the film is not completely removed.

The FWI technique has been used on devices with feature sizes down to 0.1 μm, aspect ratios up to 50:1, percent open area as low as 5%, film thickness greater than 2 μm, and substrate sizes larger than 300 mm. Endpoint, etching or deposition rate, and uniformity can be monitored for a variety of transparent thin films, including polysilicon, amorphous silicon, silicon dioxide, nitride, and photoresist. For nontransparent materials, such as aluminum, tungsten silicide, chrome, and tantalum, rate cannot be measured directly, but spatially resolved etching endpoint and thus endpoint uniformity have been determined.

Examples of the use of FWI are shown in the following figure. Figure 25.24 is an example of the signals from three different CCD pixels recorded during a polysilicon gate etching process. Each pixel imaged a small, distinct region; an image of the wafer is included to indicate the position of these pixels. Two of the pixels are on the wafer and display a periodic signal due to the change in thickness of the thin film. The pixel off the wafer shows the optical emission signal, which rises at endpoint. Analysis of the periodic

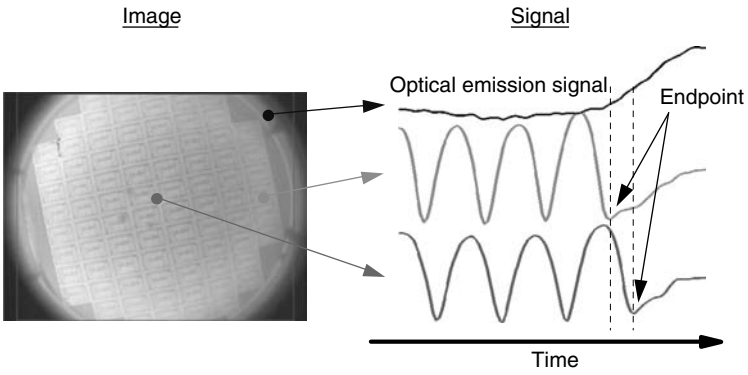


FIGURE 25.24 Signal from three positions: two on the wafer and one off the wafer. (Adapted from input by Conner, W. T., Leybold Inficon, Inc., East Syracuse, NY. http://www.inficon.com)

signal is used to determine rate and/or endpoint, while analysis of the optical emission signal is used to independently detect the average endpoint time for the entire wafer.

Figure 25.25 is an example of a full wafer etching rate uniformity surface plot. The plot was generated from rate calculations at 4000 locations on a rectangular grid covering the wafer. Two trends are evident. First, the etching rate at the center of the wafer is lower than at the edge. Second, variations within each die are visible as a regular array of peaks and valleys in the etching rate surface plot. The deepest of these valleys go all the way to zero and correspond to areas of pure photoresist mask, which did not etch appreciably in this high selectivity process.

Figure 25.26 is an example where an FWI sensor was used to automatically monitor every product wafer. Results for each wafer were determined and displayed, while the next wafer was being loaded into the processing chamber. The figure shows the etching rate and uniformity for four consecutive product wafer lots. The process was stable (no large fluctuations in rate or uniformity), but not very uniform

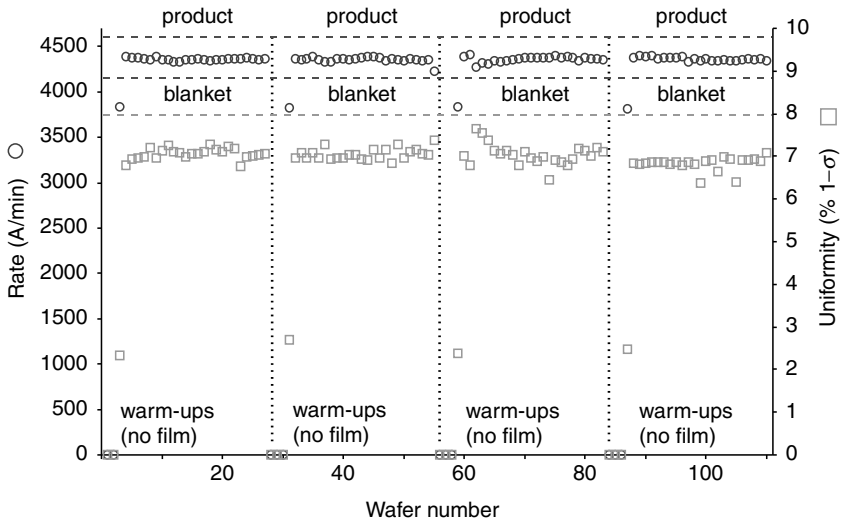


FIGURE 25.25 Full wafer etching rate map. Average = 2765 Å/min, uniformity = 3.9% $1-\sigma$. (Adapted from input by Conner, W. T., Leybold Inficon, Inc., East Syracuse, NY. http://www.inficon.com)

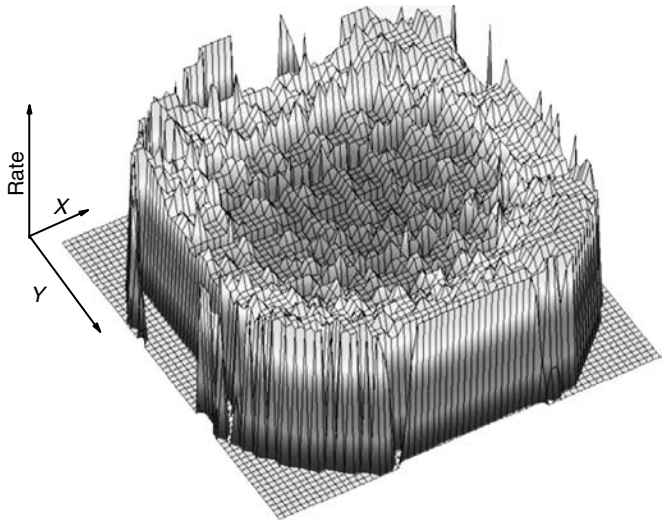


FIGURE 25.26 Rate and uniformity for four product wafer lots. (Adapted from input by Conner, W. T., Leybold Inficon, Inc., East Syracuse, NY. http://www.inficon.com)

(7%; $1-\sigma$). Furthermore, pattern-dependent etching is clearly evident. At the beginning of each lot, several bare silicon warm-up wafers and one blanket (not patterned) wafer were run, then the patterned product wafers were run. The blanket wafers etched about 10% slower and much more uniformly than the product wafers. The difference between the blanket and the product wafers demonstrates the need to use real product wafers to monitor a process.

Sensor calibration has been achieved by a comparison between FWI sensors and ex-situ film thickness metrology instruments. The agreement is generally good, even though the two systems do not measure exactly the same thing. The FWI measures dynamic changes in film thickness, while the ex-situ instruments measure static film thickness. It is typical to take the thickness—before minus thickness—after measured ex-situ and divide this by the total processing time to get the ex-situ rate and uniformity values that are compared with the rate and uniformity measured in-situ by the FWI.

Integration to the processing tool is required to obtain the benefits provided by an FWI sensor. There are two main technical issues. First, a window that provides a view of the wafer during processing is required. Between wet cleans, this window must remain transparent enough that the wafer stays visible. Second, communication between the tool's software and the FWI sensor's software is useful to identify the process and wafer/lot and to synchronize data acquisition. Both of these technical needs must be met whether the FWI runs on a separate computer or on a subsystem of the tool controller.

The FWI sensor provides different benefits to different users. In R&D, it provides immediate feedback and detailed information that speeds up process or equipment development and process transfer from tool-to-tool. In integrated circuit (IC) production, every product wafer can be monitored by the FWI sensor so that each wafer serves as a test wafer for the next. This means that fewer test, monitor, qualification, and pilot wafers are required—a significant savings in a high-volume Fab. Also, fewer wafers are destroyed before faults or excursions are detected; and data are provided for SPC of the process.

25.3.1.2 Current Sensor for Film Removal in CMP

There are two distinct measurement and control problems in CMP. The first is the measurement of the thickness of a layer during the blanket thinning of that layer by CMP. Since there is no endpoint to this process, the only way to control it is via a pre- and post-measurement of the specific layer thickness.

This has already been discussed in Section 25.3.1.1.1.3. The second is the determination of the endpoint when polishing a metal layer on an oxide substrate. This is an easier problem that has been solved by monitoring the current to the motor that rotates the wafer carrier. Most systems rotate the carrier at constant RPM. This requires the current supplied to the motor to increase or decrease depending on the amount of drag. Fortunately, this drag changes fairly reproducibly as the system polishes through the metal and reaches the oxide interface. A variety of other factors also influence the total motor current, hence there is considerable noise to this signal; however with proper signal conditioning, the endpoint can be detected. In one such endpoint system [58], the motor current signal is amplified, normalized, and high frequency components are removed by digitally filtering. Proprietary software is then used to call endpoint from the resultant trace.

Other film stacks such as poly/oxide, ploy/nitride, or oxide/nitride may find this technique useful as well but this will need to be tested and will likely require different signal conditioning and endpoint software algorithms. The motor current signal also appears to be useful to diagnose other tool parameters as well. Correlation between deviations of "known good" data traces and specific tool problems may allow the user to diagnose tool failures and to signal preventative maintenance. This sensor typically comes integrated into the tool, by the OEM supplier.

25.3.1.3 Photo-Acoustic Metrology

Section 25.3.1.1 described optical methods for the measurement of optically transparent films. There is also a need for a simple measurement of metal film thickness. This section [59] describes an impulsive stimulated thermal scattering (ISTS) method for noncontact measurement of metal film thickness in semiconductor manufacturing and process control. The method, based on an all-optical photoacoustic technique, determines thickness and uniformity of exposed or buried metal films in multilayer stacks with repeatability at the angstrom level. It can also be used to monitor CMP processes and profile thin metal films near the edge of a wafer. The method is being investigated for use in monitoring both the concentration and the depth of ions, including low-energy low-dose boron ions implanted into silicon wafers. While currently this technology is implemented in an off-line tool, it has the potential to be developed as an in-situ sensor for measuring properties of both metal films and ion-implanted wafers.

25.3.1.3.1 The Photoacoustic Measurement Technique

The photoacoustic measurement method used in this tool [60] is illustrated schematically in the inset to Figure 25.27. Two excitation laser pulses having a duration of about 500 ps are overlapped at the sample to form an optical interference pattern containing alternating "light" (constructive interference) and

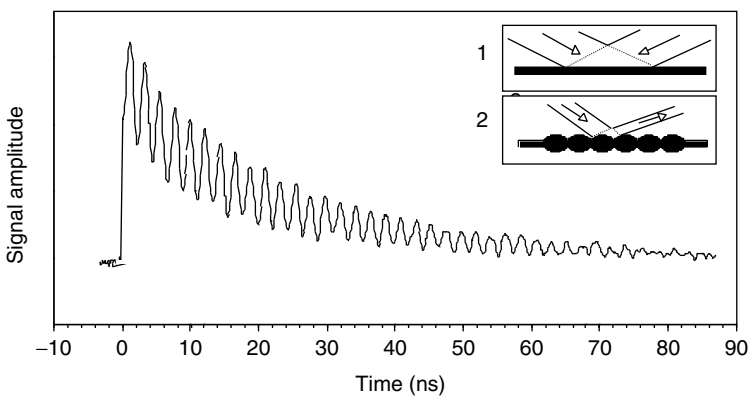


FIGURE 25.27 Signal waveform measured from a 1-mm aluminum film. (Adapted from input by Hanselman, J., Active Impulse Systems, Natick, MA.)

"dark" (destructive interference) regions. Optical absorption of radiation in the light regions leads to sudden heating and thermal expansion (Box 1 in Figure 25.27). This launches acoustic waves whose wavelength and orientation match those of the interference pattern, resulting in a time-dependent surface "ripple" that oscillates at the acoustic wave frequency [61]. A probe laser beam irradiates the surface ripple and is diffracted to form a signal beam that is modulated by the oscillating surface ripple (Box 2 in Figure 25.27). (The displacement of the surface is grossly exaggerated for purposes of illustration.) The signal beam is then detected and digitized in real time, resulting in a signal waveform such as the one in Figure 25.27. With this method, data are measured in real time with very high SNRs: the data shown were collected from a 1-μm aluminum film in about 1 s.

The acoustic wave that is excited and monitored in these measurements is a waveguide or "drumhead" mode, whose velocity is a sensitive function of the film thickness. The film thickness is calculated from the measured acoustic frequency, the spatial period of the interference pattern (i.e., the acoustic wavelength), and the mechanical properties (i.e., density, and sound velocity) of the sample. The thickness determined in this manner correlates directly to traditional techniques, such as 4-point probe measurement and scanning electron microscopy (SEM) thickness determination. Moreover, the acoustic wavelength that is excited in the film can be rapidly changed in an automated fashion. Data collected at several different acoustic wavelengths can be used to determine sample properties in addition to film thickness. In particular, thermal diffusivities and the viscoelastic properties of the sample can be measured.

A modified form of the optical technique used to determine film thickness can be used to monitor the concentration of ions implanted in semiconducting materials. In this case, the waveform of the diffracted signal depends on the concentration and energy of the implanted ions. Ion concentration and depth can be separately determined from parameters of the measured signal.

25.3.1.3.2 Hardware Configuration

The photoacoustic hardware is a small-scale optical system housed in a casting measuring approximately $50 \times 50 \times 10$ cm. The optical system uses two solid-state lasers: a Nd:YAG microchip laser generates the 500 ps excitation pulses, and a diode probe laser generates the probe beam that measures the surface ripple. A compact optical system delivers these beams to a sample with a working distance of 80 mm. The spot size for the measurement is 25×100 μm. For each laser pulse, the optical signal is converted by a fast photodetector to an electrical waveform that is digitized by a high-speed A/D converter. The digitized signal is further processed by a computer to extract the acoustic frequency and other waveform parameters. A thickness algorithm calculates the film thickness from the measured acoustic frequency, the selected acoustic wavelength, and the mechanical properties of the sample.

Integrated metrology requires in-situ or in-line monitors that can attach directly to cluster tools and monitor film properties of a wafer in, or emerging from, the process chamber. This photoacoustic measurement technology fulfills many of the requirements for such a sensor. As described above, it is compact, fast, does not require moving parts and has the long working distance necessary for optical measurement through a viewing port. While currently this technology exists as an off-line metrology tool, it is easily adaptable as an in-line sensor. With appropriate optical access to the processing chamber, it is possible to evolve this methodology to an in-situ sensor for measuring properties of both metal films and ion-implanted wafers.

25.3.1.3.3 Applications

The principle application of this technology is for the measurement of metal film thickness in single and multilayer structures. Figure 25.28 shows 49-point contour maps of a 5000 Å tungsten film deposited directly on a silicon; the map on the left was measured nondestructively with the InSite 300 in about 1 min, while the map on the right was measured destructively with a four-point probe in about 4 min. The contours of the maps are nearly identical, both showing thickness variations of about 500 Å across the surface of the film.

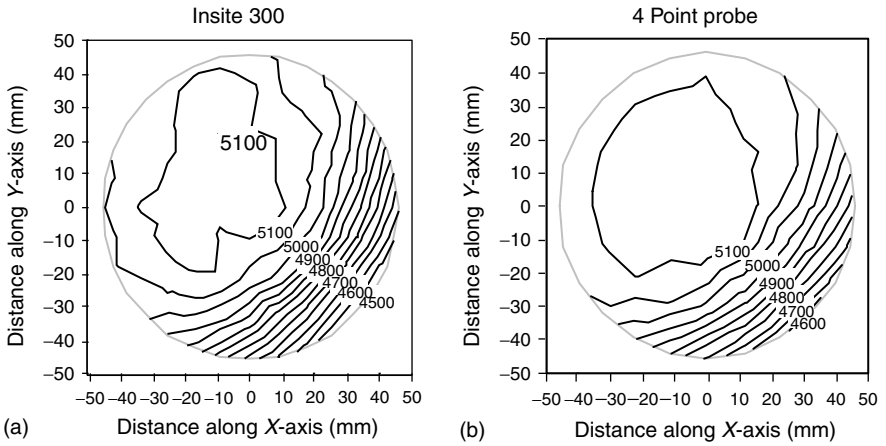


FIGURE 25.28 Comparison of 49-point contour maps of a tungsten film measured nondestructively using the Insite 300 (left) and destructively using a four-point probe (right). (Adapted from input by Hanselman, J., Active Impulse Systems, Natick, MA.)

This tool can also measure the thickness of one or more layers in a multilayer structure, such as a 1000 Å TiW film buried beneath a 2000 Å aluminum film. In this case, the system is "tuned" to explicitly measure the relatively dense buried film (TiW has a density of about 13,000 kg/m^3, compared with 2700 kg/m^3 for aluminum). This tuning is done by first initiating a low-frequency acoustic wave that is sensitive to changes in the TiW thickness, but relatively insensitive to changes in the aluminum thickness. These data are processed to generate the TiW contour map. The system then initiates a relatively high-frequency acoustic wave that is sensitive to the combined thickness changes in the TiW/aluminum structure. A contour map of the outer aluminum film can be generated from this combined data.

Full characterization of the uniformity of deposited metal films requires measurement to the edge of the film. This is particularly important for monitoring sputtering and CVD tools, which are often configured to deposit a film blanketing the entire wafer except in an "edge-exclusion zone" along the outer perimeter of the wafer. For films deposited by a process with such an edge-exclusion zone, the thickness drops from its nominal value to zero within a few hundred microns from the wafer edge. It is important to verify that edge specifications are met, as devices bordering the edge-exclusion zone can represent close to 10% of the total number of devices on a 200-mm wafer or 7% for a 300-mm wafer. Besides the contact issue, probe techniques are limited in this regard by their probe spacing and electrical issues near the edge of the wafer. The small spot size used in this methodology makes it possible to profile this narrow edge-exclusion zone.

This technique has also been applied to the measurement of the thickness of an Al film during CMP. In one particular application, 49-point contour maps were generated from data measured prior to polishing and following 30-s intervals of polishing. Prior to polishing, the film had no distinct pattern. The CMP process imparted a "bull's eye" contour to the film that is evident after about 60 s and becomes more pronounced as the polishing process continues. The data also indicate that the average removal rate is not constant, varying from ca. 60 Å/s during the first 30 s to ca. 120 Å/s in the final 30-s interval. Measurement at the center and edge of a wafer can be performed in a few seconds, making this approach attractive for real-time monitoring of CMP removal rates.

A variation of ISTS, called impulsive stimulated scattering (ISS), has been used to measure both the concentration and the energy of ions implanted in silicon wafers. An ISS is an optical technique that initiates and detects both electronic and acoustic responses in the implanted semiconductor lattice. In these measurements, the signal waveform shows features that vary with the concentration of the

implanted ions, and a separate parameter that varies with the depth of the implanted ions. Although results at this early phase are preliminary, ISS has effectively measured arsenic, phosphorous, and boron ions implanted at energies ranging from 3 keV to 3 MeV, and concentrations ranging from 1×10^{11} to 1×10^{16} cm^{-2}. The method appears to be particularly effective for measuring shallow-junction boron implants made at low energies and/or low concentrations. The ISS measures samples rapidly (less than 2 s) and remotely (12 cm working distance), making in-situ measurements a real possibility.

25.3.2 Feature Profile

Measurement for the control of lithography has classically relied on off-line metrology techniques, such as SEM, and more recently on atomic force microscopy (AFM). The SEMs are not applicable to in-situ measurements. An AFM, due to its very small field of view and slow scan rates, is also not likely to become even an in-line sensor for routine feature size measurements. Scatterometry is the only current technology that is capable of evolving into an in-line sensor for feature size measurements.

25.3.2.1 Scatterometer

25.3.2.1.1 Theory of Operation

Scatterometry, as applied in the semiconductor industry [62], is a nondestructive optical technique used to estimate wafer-state parameters, such as critical dimension, film thicknesses, and profile. The original work evolved from R&D work at the University of New Mexico [63,64], and provided estimates of wafer-state information by an analysis of light scattered, or diffracted, from a periodic sample such as resist lines in a grating. This light pattern, often referred to as a "signature," can be used to identify the shape and spatial features of the scattering structure itself. For periodic patterns, the scattered light consists of distinct diffraction orders at angular locations specified by the grating equation:

$$\sin\theta_{\mathrm{i}} + \sin\theta_n = \frac{n\lambda}{d} \tag{25.5}$$

where θ_{i} is the angle of incidence, θ_n is the angular location of the nth diffraction order, λ is the wavelength of incident light, and d is the spatial period (pitch) of the structure. The fraction of incident power diffracted into any order is very sensitive to the shape and dimensional parameters of the diffracting structure, and thus may be used to characterize that structure itself [65]. In addition to the period of the structure, which can be determined quite easily, the thickness of the photoresist, the width of the resist line, and the thicknesses of several underlying film layers can also be measured by analyzing the scatter pattern. In commercial scatterometers, the signature is generated by a variety of methods, such as varying the angle, wavelength, or polarization of the incident light [66,67]. The scatterometric analysis can best be defined in two steps. First, in what is referred to as the forward problem, the diffracted light "fingerprint" or "signature" from a periodic grating is measured using a scatterometer. As mentioned, this light can be either a single-wavelength laser beam that is varied over multiple angles, or a multiwavelength source at a fixed angle. Using the grating equation, the detector is able to track any diffraction order as the angle of incidence (or wavelength) is varied. Thus, the intensity of a particular diffraction order is measured as a function of incident angle (this is known as a scatter signature).

Figure 25.29 illustrates this technique and shows the resulting trace of the zeroth order intensity vs. the incident angle [68].

In the second step, known as the inverse problem, the scatter signature is used to determine the shape of the lines of the periodic structure which diffracts the incident light. To solve this problem, the grating shape is parameterized [69], and a parameter space is defined by allowing each grating parameter to vary over a certain range. A diffraction model, most commonly rigorous coupled-wave theory [70], is used to generate a library of scatter signatures for all combinations of parameters, and an analysis algorithm is used to compare experimental and theoretical data. The parameters of the theoretical signature that match most closely with the experimental signature are taken to be the parameters of the unknown

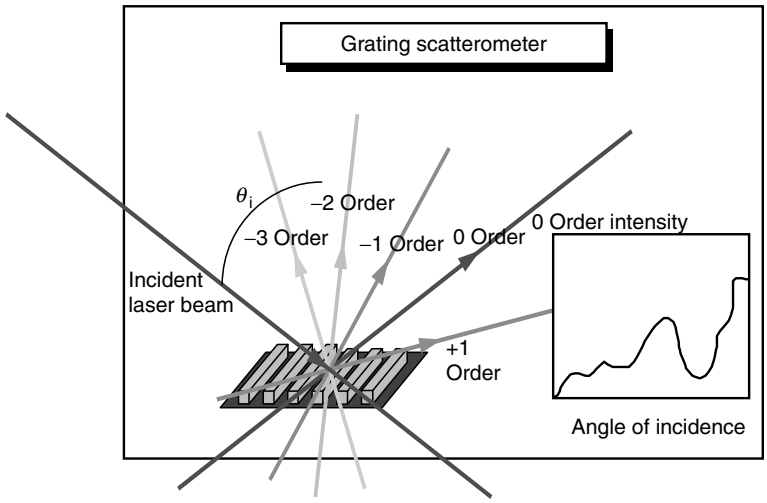


FIGURE 25.29 Diffracted orders for a grating scatterometer. (From Bushman, S., and S. Farrer, Process, Equipment, and Materials Control in Integrated Circuit Manufacturing III, Proceedings of SPIE, 3213, 79–90, 1–2 October 1997, August 1997.)

sample. One algorithm that can be used to select the closest match between theoretical and measured traces is based on minimizing a cost function such as the mean squared error, which is given by

$$\text{MSE} = \frac{\frac{1}{N}\sum_{i=0}^{N}(x_i - \hat{x}_i)^2}{\frac{1}{N}\sum_{i=0}^{N}(x_i)^2} \tag{25.6}$$

where N is the number of angle measurements, x_i is the measured reference trace, and $\hat{x}$ is the candidate trace from the theoretical library. It should be noted that because the technique relies on a theoretical model, calibration is not necessary. As an alternative to the library search method to determine the signature that minimizes the cost function, "real-time" regression can be used to estimate each parameter, although this technique is limited by the computing power used for the analysis [71].

Figure 25.30 depicts an example of an experimental signature in comparison with theoretical data, and illustrates the sensitivity of the technique for linewidth measurements. In the figure, the two theoretical scatter signatures correspond to two linewidths which differ by 10 nm. The difference between the two signatures is quite noticeable. The experimental data for this sample—a 1-μm pitch photoresist grating with nominal 0.5-μm lines—matches most closely with the 564-nm linewidth. Thus, the signatures provide a useful means for characterizing the diffracting features.

25.3.2.1.2 Applications

As a metrology technique, scatterometry provides a number of advantages over more traditional techniques of CD-SEM and AFM for characterizing feature profiles. Since scatterometery is an optical technique, it has a clear advantage in measurement time as there is no additional overhead for placing the sample in a vacuum system—as in the case of CD-SEM, or sampling the surface—as in the case of the AFM. Also, scatterometers have the ability to estimate sidewall angle information, maintain relative immunity to resist shrinkage, and provide capability to measure notched or footed profiles. However, scatterometry does require knowledge of the optical properties of additional films, in order to ensure accurate models of grating and line shape parameters [72,73].

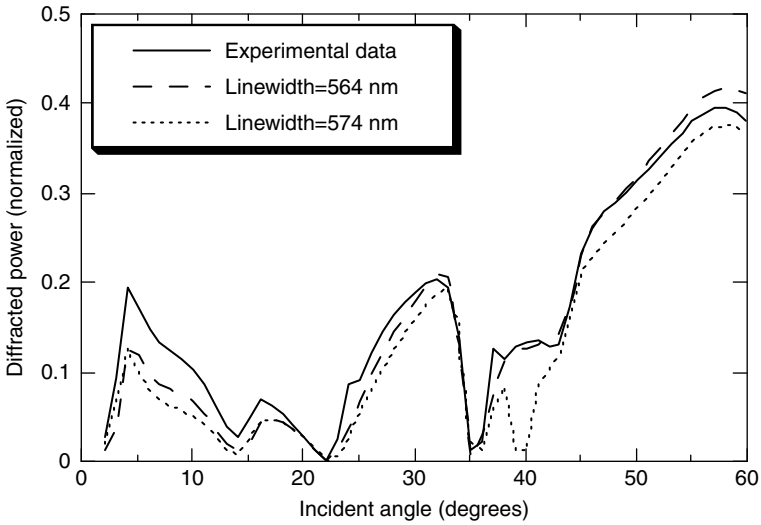


FIGURE 25.30 Scatter signatures for two linewidths differing by 10 nm. (Adapted from input by Christoper Raymond, Accent Optical Technologies, Bend, OR, http://www.accentopto.com)

Since the introduction of commercial scatterometry in 1995, the technique has been applied to a number of semiconductor industry applications [74]. Initial applications focused online/space gratings of simple resist profiles approximated as trapezoids on silicon substrates [75], and further developments of this application have led to lithography process monitoring [76]. Figure 25.31 provides a cross-sectional representation of the more complex features that are common in resist, etch, spacer, or contact gratings. To accommodate these profile shapes, multiple trapezoids—each with its own width, depth, and sidewall angle and possibly with rounded corners—can be used. The analysis of a resist grating on the polysilicon gate film stack (Figure 25.31A) or the silicon trench isolation film stack (Figure 25.31B), is more complex than the analysis of a resist grating on silicon due to the additional underlying films that now have to be optically characterized. Post-etch, the polysilicon line, and the shallow trench isolation (STI) trench are not generally modeled by single trapezoids, but require multiple trapezoids (possibly with rounded corners) to capture the profile of interest (Figure 25.31C and Figure 25.31D) [7,11,14,77]. With improvements in computing capability, more complex scatterometry applications can now be performed. Latest applications in scatterometry include the 2D offset spacer grating where a conformal film surrounds the standard polysilicon gate [78] as shown in Figure 25.31E. Even greater complexity is found in contact and metal etched arrays which are 3D in nature, as shown in the top-down view in Figure 25.31F [79–81], which require more sophisticated models that account for the additional dimensionality. Additional complexity can be added in each of these applications—if more accuracy is required—by considering additional parameterization of the profile, or by including additional film layers, as necessary. For some applications, scatterometry is one of the few methods available to estimate complex profile and film information for multilayer stacks, such as back-end trench layers [6] or memory structures [82].

As scatterometry has become more accepted in the semiconductor industry, further development has been invested in migrating scatterometry to an in-situ metrology. Current commercial examples include the addition of a scatterometer head on the litho track and into the robot handler for plasma etch [83].

25.3.2.1.3 Summary/Conclusions

In summary, scatterometry is an optical technique where raw signature data collected from the sample under test are compared with a signature estimated from a detailed physical model that includes all film

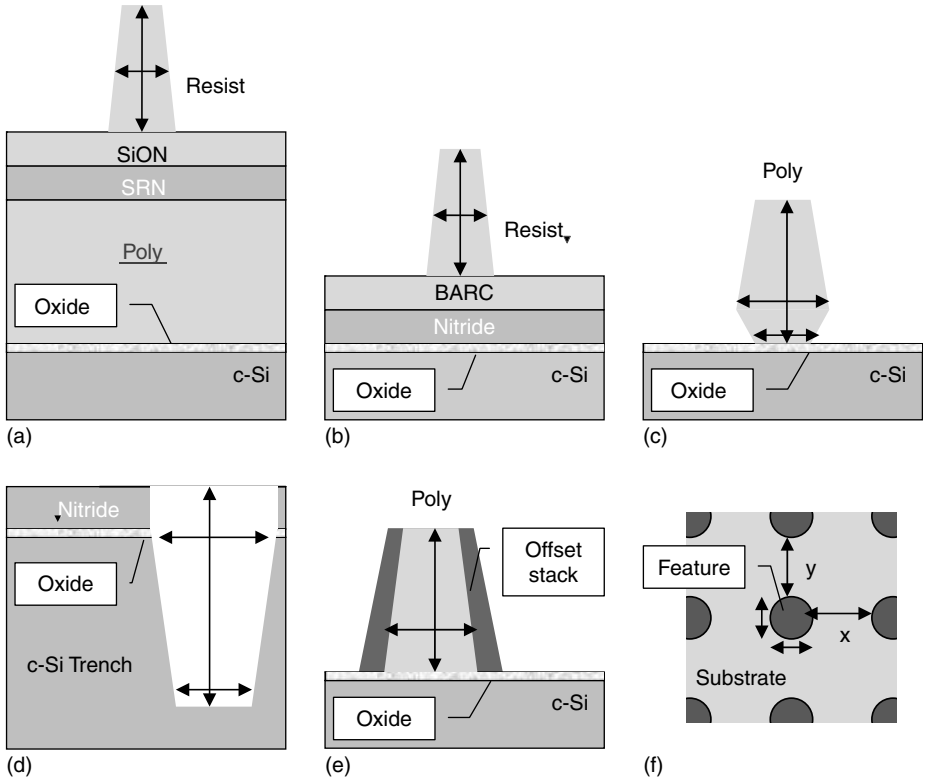


FIGURE 25.31 Graphical depiction of common scatterometry feature profiles.

and grating information. One method to solve the detailed physical model for the signature is to use a technique called rigorous coupled-wave theory [3]. As this is a computationally expensive algorithm, the common practice is to generate a discrete library of all feasible variations (pitch, CD, sidewall angle, film optical properties, etc.) for a given structure. This library of signatures is then compared with the physical signature and the parameter set of the closest fit match (in a least-squares sense) is returned as the estimate for the physical structure. Regression-based solutions are also available to improve the resolution of this technique. Numerous applications have been developed using scatterometry in the semiconductor process flow, in particular in the areas of lithography and etch process control. Recent advances in computing speed have improved the feasibility of using the technique for more complex 3D structures.

25.4 Measurement Techniques for Potential Sensors

A number of metrology tools exist based on sensing methods that are currently implemented on large, complex hardware. These methods are currently used for ex-situ measurements. However, adaptations of these techniques can become implemented as in-situ tools in OEM processing tools. This evolution will be paced by our abilities to generate less expensive and more compact versions of these tools, and by the need to implement such methods as in-situ sensors for fault detection or MBPC. Since some of these methods are likely to find their way into OEM processing tools within the next 3–5 years, a brief description of these methodologies is warranted.

25.4.1 Ellipsometry

Ellipsometry, in its single-wavelength, dual-wavelength, or spectral embodiments, is a well-established technique for the measurement of film thickness. The fundamentals of ellipsometry are described in the chapter on in-line metrology. So the following emphasizes the in-situ aspects of this sensor.

25.4.1.1 Theory of Operation

Ellipsometry is the science of the measurement of the state of polarization of light [19]. The polarization of light, in the usual conventions, corresponds to the spatial orientation of the E-field part of the electromagnetic wave. Since light is a transverse wave, the polarization is 2D; there are two independent orientations of the polarization in a vector space sense. The linear bases are usually referred to as the *P* and *S* components. For light that reflects from a surface, or is transmitted through an object, the *P* polarization lies in the plane of reflection (or transmission), and the *S* polarization is perpendicular to the plane of reflection. In addition to specifying the amount of *P*-type and *S*-type components, the phase difference between them is also important. The phase lag is a measure of the difference in the time origin of the two (*P* and *S*) electric field vibrations.

For the case where there is a nonzero phase lag the E-field vector traces out a curve in space. The projection of this curve onto a plane that is normal to the direction of propagation of the light is generally an ellipse, thus the origin of the name "ellipsometry." A special case of the ellipse is a circle. There are two unique circular polarizations referred to as left-handed and right-handed depending on whether the *P*-polarization component leads or lags the S-polarization component. At times it is convenient to use the circular basis set in place of the linear basis set. The full state of polarization of light requires the specification of a coordinate system and four numbers, which includes the amount of unpolarized light. Unpolarized light implies that the polarization is randomly and rapidly changing in time. This naturally occurs for a light source that consists of a large number of independent, random emitters.

The reflectivity of light from a surface depends on the state of incident polarization. For example at a specific angle for many materials, the *P*-polarization has no reflection, whereas the S-polarization does. The angle at which this occurs is the Brewster angle of the material, it is a function of the index of refraction of the material. Using fully polarized light, the index of refraction of a bulk material may be readily measured by finding the Brewster angle.

Figure 25.32 illustrates the difference in reflection for *P* and *S* polarizations as a function of incidence angle for 0.5 μm light (green light). The Brewster angle at about 76° is the point where the *P* polarization

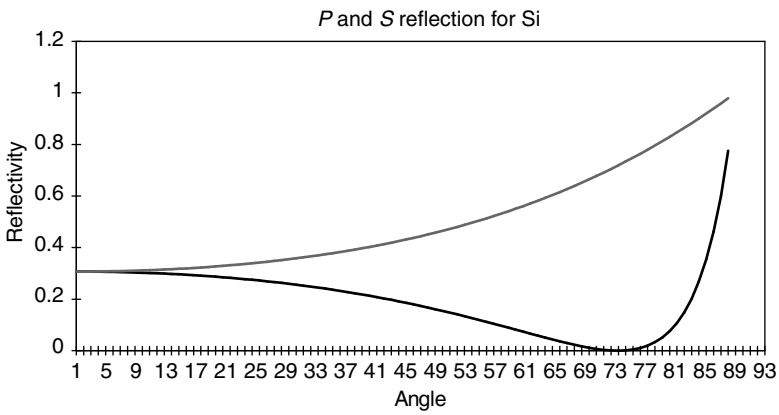


FIGURE 25.32 Reflection from silicon for 0.5 mm light for *P* and *S* polarizations. (Adapted from input by Whelan, M., Verity Instruments, Carrollton, TX. http://www.verityinst.com)

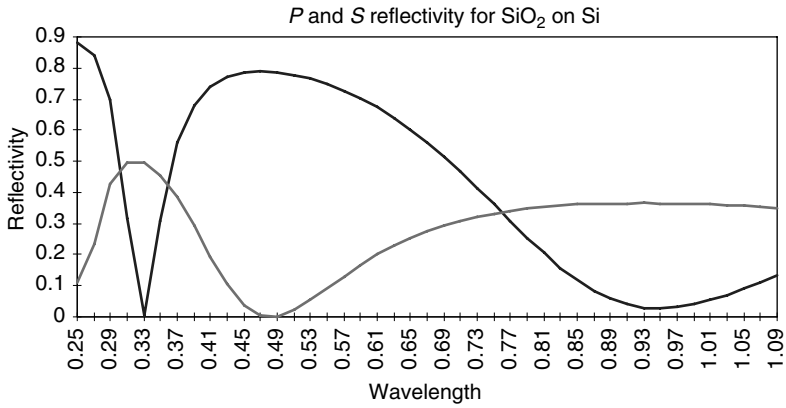


FIGURE 25.33 Reflection from SiO_2 layer on Si for *P* and *S* polarization's as a function of wavelength. (Adapted from input by Whelan, M., Verity Instruments, Carrollton, TX. http://www.verityinst.com)

reflectivity goes to 0. At normal incidence and grazing incidence, there is generally much less information about the surface that can be derived from measurements in the change of polarization.

By taking ratios, all the information about the change in the state of polarization may be determined by specifying two numbers. These are known as the ellipsometric parameters ψ and Δ. Psi (ψ) is usually given as the angle whose tangent is the ratio of the magnitudes of the *P* and *S* components of the reflected light. Delta (Δ) is the relative phase between the *P* and *S* components. By measuring ψ and Δ as a function of incident angle, or as a function of wavelength at a fixed incident angle, much information may be determined about the reflecting surface including the effect of thin films (thickness and compositions). Figure 25.33 shows the reflectivity of the *P* and *S* polarizations from a sample of silicon dioxide (220-nm thick) layer on silicon substrate, as a function of wavelength. This example illustrates the type of data from which the thin film thickness and material composition is inferred by regressing on the ellipsometric equation that are defined by this system.

25.4.1.2 System Integration

When the wavelength of the incident light varies (using a white light source) for a fixed incident angle the term spectral ellipsometry is used. For a fixed wavelength (laser source) with variable incident angle the term variable angle ellipsometry is used. Instruments that vary both angle and wavelength are variable angle spectral ellipsometers.

Figure 25.34 is a schematic representation of a spectral ellipsometer. The white light source is a broadband emitter such as a Xenon arch discharge. The fixed polarizer passes a specific linear polarization component. The polarization modulator is a device that changes the polarization in a known manner such as a photo-elastic polarization modulator. In some instruments, the function of these two devices is replaced with a polarization element that is mechanically rotated. The analyzer selects out a specific state of polarization of the reflected light. Since the state of the incident polarization is well defined, the effects of reflection from the sample can be determined. The spectrograph analyzes the white light source into a number of spectral components.

The use of spectral ellipsometers for in-situ monitoring and control has been limited by the cost of these units. An additional constraint has been the complexity of the integration of these optics (two opposing windows required) into standard OEM processing tools. The cost issue is slowly improving through the development of lower-cost ellipsometers. When warranted, the optical complexity can be overcome [84]. As processing complexity and the inherent cost of misprocessing 200–450-mm wafers continues to increase, spectral ellipsometers will likely find their way into OEM tools for in-situ monitoring and control of thin film growth and composition in real time.

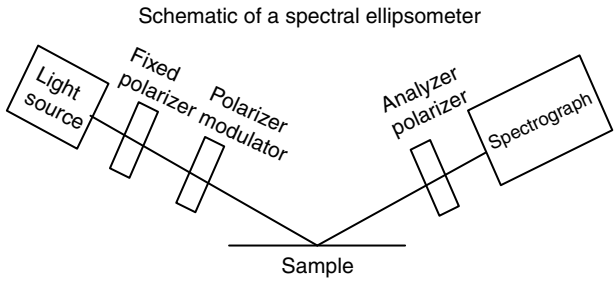


FIGURE 25.34 Schematic representation of a spectral ellipsometer. (Adapted from input by Whelan, M., Verity Instruments, Carrollton, TX. http://www.verityinst.com)

25.4.2 Epi Resistivity and Thickness

In the process of depositing an epi layer of silicon, resistivity, and thickness of the epi layer are the two parameters of greatest interest [85]. Traditional methods of monitoring epi layer resistivity measures either the average resistivity of the layer, as is the case of a four-point probe, or the resistivity as a function of depth into the epi layer, as is the case with a Hg probe or CV Schottky diode. These traditional methods are all destructive, as probe marks, contamination due to Hg, and metal dots all contaminate the wafer.

A new technique has recently been developed [86] that performs a nondestructive measurement of epi layer resistivity and profile. The technique used is conceptually quite similar to Hg probe or CV Schottky measurements. While the technique was introduced as a stand-alone metrology tool, an in-line version incorporated into the cooling station of an epi reactor is an obvious extension of the technology.

25.4.2.1 Theory of Operation

Both the CV Schottky diode and Hg probe techniques place an electrode on the surface of the semiconductor and then measure the depletion width by looking at the capacitance across the depletion width. They vary the depletion width by varying the voltage on the electrode and measure the capacitance of the depletion width at each electrode voltage. Similarly, this technique positions an electrode near the semiconductor surface, although in this case it does not touch the wafer. It then measures the depletion width for each of multiple voltages on the electrode.

The technique used to position the electrode near the semiconductor surface, but not touching it, is similar to the air-bearing effect used in computer hard disk drives, and is shown in Figure 25.35. A disk

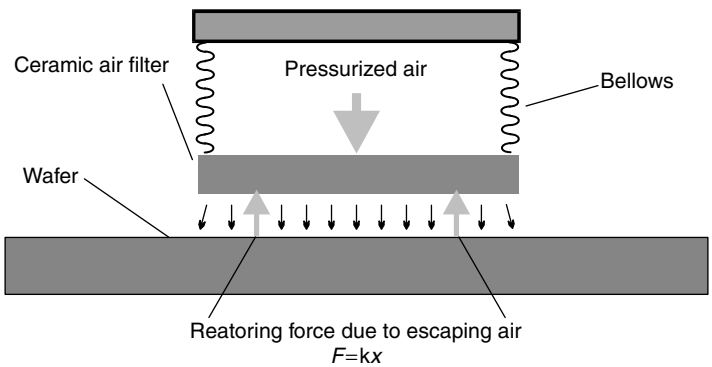


FIGURE 25.35 Air-bearing mechanism. (Adapted from input by Charlie Kohn, SemiTest, Inc., Billerica, MA.)

whose bottom surface is made of porous, inert material is pushed toward the wafer surface by air pressure above the porous surface. As air escapes through the porous surface, a cushion of air forms on the wafer, and the air cushion acts like a spring and prevents the porous surface from touching the wafer. The porosity and air pressure are designed such that the disk floats approximately 2 μm above the wafer surface. A stainless steel bellows acts to constrain the pressurized air and to raise the porous disk when the air pressure is reduced.

Note that if the air pressure fails the disk moves up, rather than falling down, and damaging the wafer. Similarly, an electrical failure would not damage the wafer surface. The mechanism is simple, as no electrical or computer feedback of any kind is required. It is analogous to suspending an object between two springs of different spring constants. The porous disk has a hole in the center, and a sensor element is mounted in the hole to prevent the pressurized air from escaping. The sensor consists of an electrode that is 1 mm in diameter. The electrode is made of a material that is electrically conductive and optically transparent. The electrode protrudes from the bottom of the porous disk, such that during the measurement it is located about one-half micron above the wafer surface. A block diagram of the measurement system is shown in Figure 25.36.

As with Hg probe and CV Schottky measurements, depletion width is measured by looking at the capacitance of the depletion layer. The system actually measures the capacitance from the wafer chuck to the electrode, which is the series combination of three capacitances; the capacitance from the wafer chuck to the wafer, in series with the capacitance of the depletion layer, and in series with the capacitance of the air gap. The capacitance of the wafer chuck to the wafer can be ignored, as the area of the wafer is so much larger than the area of the electrode. Even with a 6-in. wafer, the ratio of the areas is more than 22,000, and although the effective separations of the capacitor plates may be unequal, it is reasonable to consider the wafer chuck to wafer capacitance as a short circuit. The capacitance of the air gap cannot be treated so easily, but because there is some electrode voltage at which the semiconductor surface is in accumulation and the capacitance of the depletion width is infinite, the capacitance of the air gap can be measured. Assuming that the actual air gap does not vary with changing electrode voltage, the capacitance of the air gap is the measured capacitance at its maximum value. Subtracting the capacitance of the air gap from

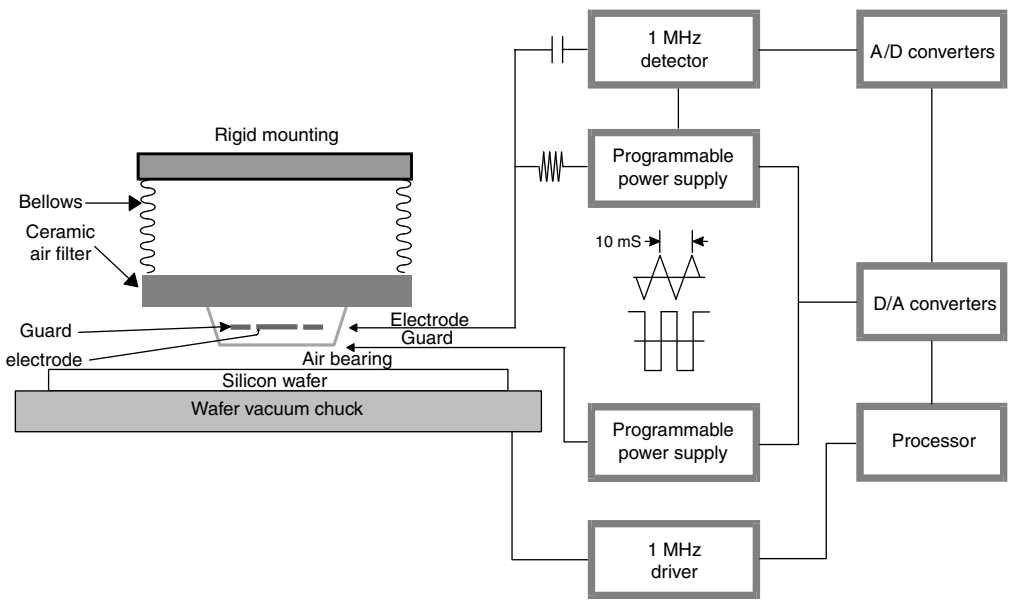


FIGURE 25.36 Block diagram of complete system. (Adapted from input by Charlie Kohn, SemiTest, Inc., Billerica, MA.)

the measured capacitance provides the capacitance of the depletion width. If the air bearing does not have infinite stiffness and the air gap changes as a result of the varying electrostatic attractive force created during the measurement, then it is possible to model the behavior and calculate the air gap capacitance at any electrode voltage.

At every step in electrode voltage, the capacitance is measured and the charge on the electrode is calculated as the integral of *C*d*V*. The relevant equation necessary to compute the profile of N_{sc} as a function of depth, *W*, are as follows

$$W = \varepsilon_s \varepsilon_0 A \left[\frac{1}{C_{total}} - \frac{1}{C_{air}} \right]$$
$$dQ = C_{meas} dV \tag{25.7}$$
$$N_{sc}(W) = \frac{dQ}{qA\, dW}$$

where A is the area of the electrode, ε refers to dielectric constant, and q is the elementary charge.

Unlike in traditional Hg probe or CV Schottky measurements, the electrode voltage in this system varies rapidly. A full sweep from accumulation to deep depletion is done in about 10 ms, and data from multiple sweeps are averaged in order to reduce the effect of noise. The fast sweep period also serves to reduce inaccuracies due to interface states and carrier generation.

The system displays either plots of resistivity vs. depth or N_{sc} vs. depth. Resistivity is obtained by converting as per the American society for testing and materials (ASTM) standard. A typical profile produced by the system is shown in Figure 25.37. Repeatability and reproducibility are quite reasonable compared with other techniques. Resistivity of a single wafer measured at 8-h intervals over a 3-day period showed a measurement error of 0.75% $(1-\sigma)$.

25.4.2.2 Calibration and Performance Range

The system is calibrated by presenting it with one or more wafers of known resistivity. The system then creates a piecewise linear calibration curve so that there is complete agreement at the calibration points and between calibration points the system interpolates to obtain good calibration. The more calibration points there are, the better the performance across the entire range of calibrated values.

Doping concentration profile can be generated within depletion depth. Figure 25.38 shows the maximum epi layer depth (for *p*-type silicon) which the system can deplete. As with mercury probe or

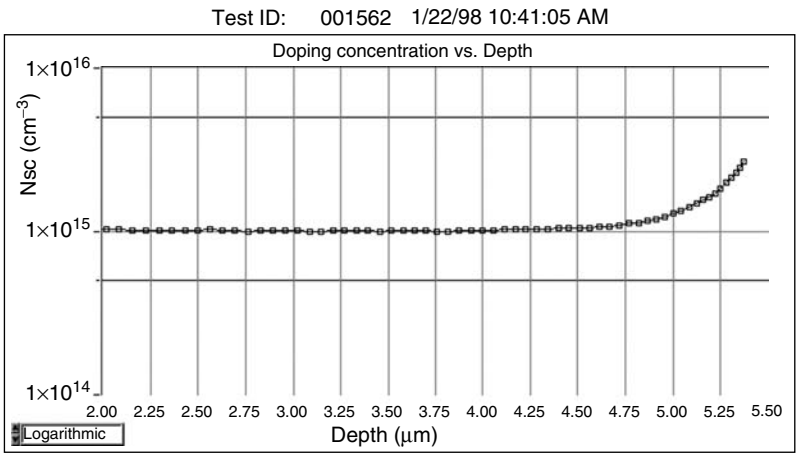


FIGURE 25.37 Epimet measurement profile. (Adapted from input by Charlie Kohn, SemiTest, Inc., Billerica, MA.)

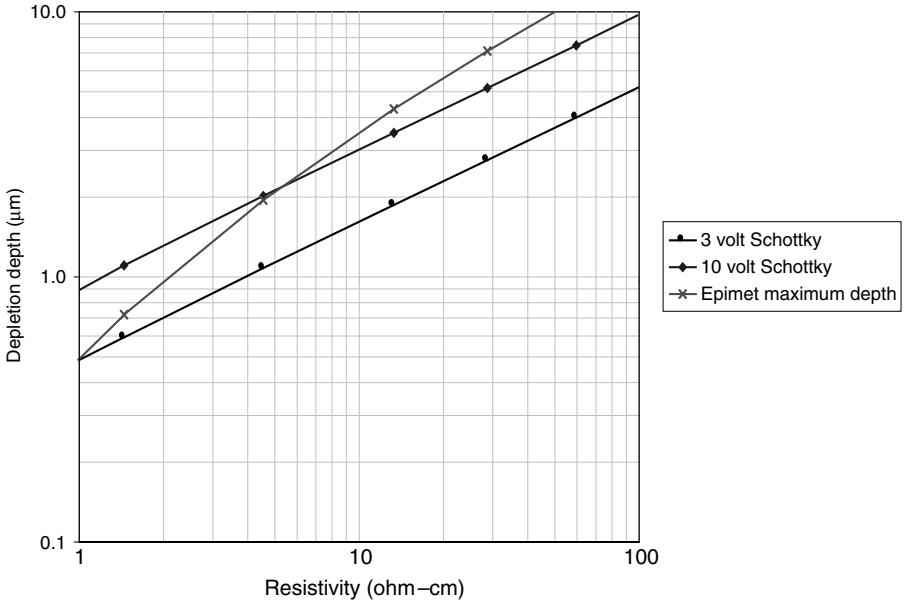


FIGURE 25.38 Epimet operating range, *p*-type epi. (Adapted from input by Charlie Kohn, SemiTest, Inc., Billerica, MA.)

CV Schottky diode, the maximum depletion is a function of resistivity, for a given applied voltage. For reference, Figure 25.38 also shows the maximum depth which can be measured by mercury probe using 3 and 10 V bias voltages. Development is underway to increase the maximum layer depth, which can be measured, i.e., move the line "up" on the graph.

25.5 Software for In-Situ Metrology

In-situ sensors in OEM tools are an absolute prerequisite for providing the benefits of the APC paradigm to SC manufacturing operations. But the sensors themselves are just one part of the system required to generate the necessary information for APC. Extensive software is required to turn the sensor data into useful information for APC decisions. Software is required for data collection, data analysis for FDC and the controllers that perform MBPC.

25.5.1 Data Collection Software

The two major sources of data used in APC applications are signals from the processing tool and from add-on sensors connected to the tool. The former is generally collected through the Semiconductor Equipment Communications Standard (SECS) interface available on the tool. The SECS protocol enables the user to configure bi-directional communications between tools and data collection systems. This standard is a means for independent manufacturers to produce equipment and/or hosts, which can be connected without requiring specific knowledge of each other. There are two components to SECS. The SEMI Equipment Communications Standard E4 (SECS-I) defines the physical communication interface for the exchange of messages between semiconductor processing equipment (manufacturing, metrology, assembly, and packaging) and a host, which is a computer or network of computers. This standard describes the physical connector, signal levels, data rate, and logical protocols required to exchange messages between the host and the equipment over a serial point-to-point data path. This standard does

not define the data contained within a message. The second component is the software standard such as SEMI Equipment Communications Standard E5 (SECS-II), that is a message content standard that determines the meaning of the messages. While SECS-I has solved the hardware interface issues, SECS-II implementation contains enough tool-specific latitude to make interfacing to individual tools a time-consuming task. The genetic model for communications and control of manufacturing equipment (GEM) standard, built upon the SEMI E5 (SECS-II) standard, specifies the behavior of a generic equipment model which contains a minimal set of basic functions that any type of semiconductor manufacturing equipment should support. The SECS-II standard provides the definition of messages and related data items exchanged between host and equipment. The GEM standard defines which SECS-II messages should be used, in what situations, and what the resulting activity should be. Brookside software [87] is a commonly used package for obtaining machine data from the SECS port on plasma etchers.

Add-on sensors are most often sophisticated enough (OES, RGA…) that they are run by a dedicated PC. This PC also collects the raw sensor data and transfers it to the analysis software package.

25.5.2 FDC Analysis Software

Once the machine and sensor data are properly formatted and available on a routine basis, FDC is achieved by the univariate or multivariate analysis of individual or multiple signals, respectively. For plasma etching, univariate analysis has been performed on the endpoint signal for a long time [88,89], primarily due to its simplicity and effectiveness in detecting processing faults. The basis of these analyses is to examine the shape of a single signal, and use some algorithm to detect a significant variation between the signal from the currently analyzed wafer relative to an accepted "reference trace" from a good wafer. In the mid-1990s, multivariate FDC became feasible, enabled by a number of software vendors [90–96] with capabilities for analyzing machine and sensor data. The capabilities of these vendors to provide multivariate FDC was evaluated in a SEMATECH FDC Benchmarking study [97]. These analyses determine the correlation between the various time-series signals coming from the tool and associated sensors. Models are generated from a number of "good wafers" that represent the normal variability of the individual signals. The major issues are the choice of the signals, the choice of the "good wafers," and the methods used for the analysis of this large volume of time-series data. An important feature of any such multivariate FDC technique is that the method be robust to the long-term steady drift in the sensors signals (that are not associated with processing faults), while it stays sensitive to small but significant signal variations at any particular time [98]. The availability of pertinent information from the factory computer-integrated manufacturing (CIM) system (e.g., lot number, wafer number, log point) is a critical component for both univariate and multivariate FDC; as these form the basis for sorting the machine and sensor signals into the necessary groups for analysis. This analysis can be performed on a local computer adjacent to the tool or on the factory network. A major driver for determining where to run the analysis is the network latency period; real-time analyses are generally performed on a local computer, while wafer-to-wafer or lot-to-lot analyses can be performed on the CIM network.

25.5.3 Model-Based Process Control Software

The MBPC is based on models of the wafer-state properties as a function of the input parameters. With such models, and routinely available feed-forward or feedback data from the factory CIM system, MBPC can be performed to keep the output parameters of interest under tight control. The numerical basis for this can be quite straightforward, and is now enabled by commercial software that performs these calculations [99,100]. Full automation of such methods, where the MBPC software obtains the necessary information from factory automation and downloads the new recipe to the tool, is a complex and generally lengthy integration task. Such integration is generally reserved for situations where benefits of MBPC have been proven on a local, off-line basis.

25.6 Use of In-Situ Metrology in SC Manufacturing

Semiconductor manufacturing has historically relied on SPC for maintaining processes within prescribed specification limits. This passive activity, where action is taken only after a process variation limit has been exceeded, is no longer adequate as device geometries shrink into the 0.25 μm range and beyond. Active control of the process is required to keep the wafer properties within the ever-decreasing specification limits. In order to achieve this tighter control, a new paradigm called APC is emerging. All APC activities are predicated on the timely observability of the process. This is the major driving force for the implementation of in-situ sensors in OEM processing tools. These sensors determine the process, wafer or machine states during the sequential processing of wafers and hence provide the necessary information for APC. The two major components of APC are FDC and MBPC. The operational benefits of APC, which are the main drivers for this operating paradigm, are:

1. *Fault detection, classification*: detect and analyze faults for enhanced yield and faster tool repair.
2. *Fault interdiction*: eliminate the continuing misprocessing of wafers.
3. *Fault prognosis*: convert from scheduled to preventive maintenance to reduce future misprocessing, resulting in higher equipment availability.
4. *MBPC*: increase yield by keeping process on target, reduce pilot wafer usage.

25.6.1 Fault Detection and Classification

An FDC determines anomalous processing conditions by univariate or multivariate (single or multiple signals) analysis methods. These conditions can be intermittent in nature, or can result from a gradual drift in the processing tool. The first task [101] of FDC is *fault detection*, i.e., determining that during the processing of a particular wafer, the sensor signatures indicate a "non-normal" state. This requires a model to be generated that represents the normal process states (sensor signals) of the tool. This is a bigger problem than might first be envisioned, since the "normal" state is not a stationary point, but a slowly changing trajectory through time. So a major requirement of these FDC methods is that they are robust against the normal drifts in the system for extended time periods, yet stay sensitive to small excursions at any point in time. The primary reason for this is that the use of such methodology in a manufacturing operation requires that there is minimal model "upkeep," adequate sensitivity to errors but a minimal number of false alarms. Data from successive wafers are then analyzed against this model and the appropriate statistics indicates whether the wafer was processed in the "normal" way or that some aspect of the process was anomalous.

Once a fault is detected, the next task is *fault classification*. Depending on the FDC algorithm, the model can be used in a "reverse" sense to establish what signal generated the fault. Even if not very specific (e.g., fault was due to a pressure error), such information is very valuable in narrowing the focus of R&M by isolating which variables causing the fault.

Note that to this point, the only indication generated by the FDC system is that the conditions during the processing of a given wafer were anomalous from the "normal" state. This does not necessarily indicate that the wafer has been misprocessed (although there is a good reason to suspect that). Even without having this final link between the process anomalies and the resulting wafer conditions, FDC can be quite valuable particularly to Equipment Engineers, who are tasked with keeping tools operating in a consistent manner. For them, knowing that a tool is behaving in an anomalous fashion is key; and their task is much simplified if the FDC method points them to the source of the problem. However, the FDC method is only truly complete when the correlation is established between the various types of faults and wafer parametrics (e.g., yield, defects). In many cases, this is a difficult task if for no other reason than the time interval between a specific process (especially towards the beginning of the process flow) and the parametric testing that occurs towards the end of the flow. These data are also generally in different databases, making correlation difficult. The SC manufacturers are currently instituting "data warehousing," which will facilitate the correlation between machine state and wafer properties. Once this task

is completed, then the FDC system becomes a very effective early-warning system for faults, and can have substantial financial impact in routine manufacturing.

25.6.2 Fault Interdiction

The simplest, most practical and beneficial use of in-situ metrology and fault detection analysis in SC manufacturing is for interdiction to the tool controller in the case of anomalous processing conditions. All processing tools monitor and control their setpoint values and shut themselves off if these are not attained. However, they can be totally insensitive to problems caused by the uncontrolled variables (e.g., wafers placed into an etcher with no resist, leaks in the gas lines). But problems from such sources can, in many cases, be detected from the analysis of a single sensor signal. The most notable example is the plasma emission signal in any plasma process. There are very few problems that do not have a noticeable affect on some portion of this signal. Hence, any method that analyzes the form of this signal can readily detect anomalies in the process. This analysis can be done in real time (during the processing of the wafer) or after the process has been completed. A statistically significant deviation from the expected signal can be used to shut off the reactor either in real time or before the next wafer is loaded into the process chamber. This ensures that no more wafers will be exposed to the anomalous processing conditions. In other tools, the detection of subtle faults might require the analysis of multiple signals with some multivariate analysis algorithm. With the approach of 12-in. wafers, preventing a full boat of wafers from misprocessing can easily represent a $250K saving.

25.6.3 Fault Prognosis

In addition to handling intermittent problems by fault interdiction, there is a need to handle problems due to slow changes in the manufacturing tool. Significant operational benefits can be had if certain faults can be predicted before they occur, and are prevented by scheduled maintenance. This has several benefits: (1) the maintenance can be done when the tool is idle (increasing tool availability) and (2) the gradually developed problem can be fixed before it contributes to wafer misprocessing (increasing yield). Typical of this kind of problem is the wear of electrodes in plasma systems, the degradation of pump capacity, etc. If the appropriate metric that is a good indicator of this problem is obtained from the tool controller or an add-on sensor, the value of this metric can be tracked. When it approaches limits that have previously been identified to represent an unacceptable situation, the tool can be scheduled for R&M. A good example is the tracking of a throttle valve signal in any vacuum processor. As the pump degrades, the throttle valve will gradually open more and more to maintain a given pressure at a given flow (i.e., for a given recipe). As this signal gets close to the "100% open" level, it is clear that the pump has to be serviced.

25.6.4 Model-Based Process Control

The goal of MBPC is to keep the output of any process close to a target value by manipulating the control "knobs" of the process. This is achieved by obtaining data from wafer-state sensors, where possible. If these are unavailable, "virtual sensors" can be generated from process state measurements using models that relate the process state characteristics to the wafer state of interest. This, of course, is a significant complication, to be attacked only when necessary. Once the real, or imputed, wafer state is routinely obtained, the sequential data are analyzed by SPC methods as suggested in Figure 25.1. When a statistically significant deviation is reached, process models are tuned and a new set of process conditions are calculated. Process models are typically simple linear (thickness = rate × time + c) or more complex polynomial models that represent the relationship between the input/output parameters. The tuner is a methodology for updating these models. In this way, the variance is transferred from the very costly process output to the less-expensive process input. This topic is addressed much more thoroughly in Chapter 23.

References

1. *National Technology Roadmap for Semiconductors*, 183, 1997.
2. Adapted from input by C. Schietinger, Luxtron Corporation, Santa Clara, CA.
3. Roozeboom, F. *Advances in Rapid Thermal and Integrated Processing NATO Series*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1996.
4. Schietinger, C., and B. Adams. "A Review of Wafer Temperature Measurement Using Optical Fibers and Ripple Pyrometry." *RTP '97*, New Orleans, LA. September 3–5, 1997.
5. Schietinger, C., and E. Jensen. "Wafer Temperature Measurements: Status Utilizing Optical Fibers." *Mater. Res. Soc. Symp. Proc.* 429 (1996): 289–90.
6. DeWitt, D. P., and G. D. Nutter. *Theory and Practice of Radiation Thermometry*. New York: Wiley, 1988.
7. Schietinger, C., B. Adams, and C. Yarling. "Ripple Technique: A Novel Non-Contact Wafer Emissivity and Temperature Method for RTP." *Mater. Res. Soc. Symp. Proc.* 224 (1991): 23–31.
8. Adapted from input by Jim Booth, Thermionics Northwest, Port Townsend, WA, http://www.thermionics.com
9. Varshni, Y. P. *Physica* 34 (1967): 149.
10. Hellman, E. S., and J. S. Harris Jr. *J. Cryst. Growth* 81 (1986): 38.
11. Weilmeier, M. K., K. M. Colbow, T. Tiedje, T. van Burren, and L. Xu. *Can. J. Phys.* 69 (1991): 422.
12. Johnson, S. R., C. Lavoie, T. Tiedje, and J. A. MacKenzie. *J. Vac. Sci. Technol.* B11 (1993): 1007.
13. Johnson, S. R., C. Lavoie, E. Nodwell, M. K. Nissen, T. Tiedje, and J. A. MacKenzie. *J. Vac. Sci. Technol.* B12 (1994): 1225.
14. DRS 1000™ Temperature Monitor, Thermionics Northwest, Inc., Port Townsend, WA, http://www.thermionics.com
15. Wang, Z., S. L. Kwan, T. P. Pearsall, J. L. Booth, B. T. Beard, and S. R. Johnson. *J. Vac. Sci. Technol.* B15 (1997): 116.
16. Booth, J. L., B. T. Beard, J. E. Stevens, M. G. Blain, and T. L. Meisenheimer. *J. Vac. Sci. Technol.* A14 (1996): 2356.
17. Adapted from input by Sensys Instruments, Sunnyvale, CA.
18. WTS 100 Wafer Temperature Sensor, Sensys Instruments, Sunnyvale, CA.
19. Adapted from input by Mike Whelan, Verity Instruments, Carrollton, TX, http://www.verityinst.com
20. Model SD1024 Smart Detector Spectrograph, Verity Instruments, Carrollton, TX, http://www.verityinst.com
21. Model 1015 DS, Luxtron Corporation, Santa Clara, CA, http://www.luxtron.com
22. Courtesy of Bob Fry, Cetac Technologies, Inc., Omaha, NE.
23. EP-2000 Spectrometer, Cetac Technologies, Inc., Omaha, NE.
24. PC1000 Miniature Fiber Optic Spectrometer, Ocean Optics, Inc., Dunedin, FL, http://www.oceanoptics.com/homepage.asp
25. Princeton Instruments, Inc., Trenton, NJ, http://www.prinst.com
26. Spectral Instruments, Tucson, Arizona, http://www.specinst.com
27. Mozumder, P. K., and G. G. Barna. "Statistical Feedback Control of a Plasma Etch Process." *IEEE Trans. Semicond. Manuf.* 7, no. 1 (1994): 1.
28. Neural Endpointer, Verity Instruments, Carrollton, TX, http://www.verityinst.com
29. Wise, B. M., and N. Gallagher. "PLS_Toolbox." Eigenvector Technologies, Manson, WA.
30. Adapted from input by Peter Solomon, On-Line Technologies, East Hartford, CT.
31. On-Line 2100 FT-IR spectromenter, On-Line Technologies, East Hartford, CT.
32. Solomon, P. R, P. A. Rosenthal, C. M. Nelson, M. L. Spartz, J. Mott, R. Mundt, and A. Perry. "A Fault Detection System Employing FT-IR Exhaust Gas Monitoring." Presented at *SEMATECH Advanced Process and Equipment Control Program, Proceedings* (*Supplement*), 290–294. Lake Tahoe, Nevada, September 20–24, 1997.
33. Adapted from input by R. J. Ferran and S. Boumsellek. Ferran Scientific, Inc., San Diego, CA, http://www.ferran.com

34. Transpector XPR, Leybold Inficon, East Syracuse, NY, http://www.inficon.com
35. Micropole™ Sensor System, Ferran Scientific, San Diego, CA, http://www.ferran.com/main.html
36. Ferran, R. J., and S. Boumsellek. "High-Pressure Effects in Miniature Arrays of Quadrupole Analyzers for Residual Gas Analysis from 10^{-9} to 10^{-2} Torr." *J. Vac. Sci. Technol.* A14 (1996): 1258.
37. Adapted from input by C. A. Gogol, Leybold Inficon, East Syracuse, NY, http://www.inficon.com
38. Tallman, C. A. "Acoustic Gas Analyzer." *ISA Trans.* 17, no. 1 (1977): 97–104.
39. Wajid, A., C. Gogol, C. Hurd, M. Hetzel, A. Spina, R. Lum, M. McDonald, and R. J. Capic. "A High-Speed High-Sensitivity Acoustic Cell for In-Line Continuous Monitoring of MOCVD Precursor Gasses." *J. Cryst. Growth* 170 (1997): 237–41.
40. Stagg, J. P. "Reagent Concentration Measurements in Metal Organic Vapour Phase Epitaxy (MOVPE) Using an Ultrasonic Cell." *Chemtronics* 3 (1988): 44–9.
41. Leybold Inficon's "Composer," Acoustic Gas Composition Controller.
42. Adapted from input by Kevin S. Gerrish, ENI Technology, Inc., Rochester, NY.
43. Bird Electronic Corporation, Solon, OH.
44. Applied Energy Systems, Inc., Malvern, PA, http://www.aenergysys.com
45. ENI Technology, Inc., Rochester, NY.
46. Fourth State Technology, Inc., Austin, TX.
47. Comdel, Inc., Gloucester, MA.
48. Advanced Energy Industries, Inc., Fort Collins, CO.
49. Buck, D. Texas Instruments, personal communication.
50. CSS 100 Chamber-Wall State Sensor, Sensys Instruments, Sunnyvale, CA.
51. Adapted from input by Ran Kipper, Nova Measuring Instruments Ltd., Weizman Scientific Park, Rehovoth, Israel.
52. NovaScan, Nova Measuring Instruments Ltd., Weizman Scientific Park, Rehovoth, Israel.
53. Model 2100 Process FT-IR, On-Line Technologies, East Hartford, CT.
54. Buffeteau, T., and B. Desbat. *Appl. Spectrosc.* 43, no. 6 (1989): 1027–32.
55. Abeles, F. *Advanced Optical Techniques*. Vol. 143. North-Holland, Amsterdam, 1967, chap. 5.
56. Adapted from input by William T. Conner, Leybold Inficon, Inc., East Syracuse, NY, http://www.inficon.com
57. Dalton, T., W. T. Conner, and H. Sawin. *JECS* 141 (1994): 1893.
58. Model 2450 Endpoint Controller, Luxtron Corporation, Santa Clara, CA, http://www.luxtron.com/index.html
59. Adapted from input by John Hanselman, Active Impulse Systems, Natick, MA.
60. InSite 300, Active Impulse Systems, Natick, MA.
61. Rogers, J. A., M. Fuchs, M. J. Banet, J. B. Hanselman, R. Logan, and K. A. Nelson. *Appl. phys. Lett.* 71, no. 2, (1997).
62. Adapted from input by Christoper Raymond, Accent Optical Technologies, Bend, OR, http://www.accentopto.com
63. McNeil, J. R., S. Naqvi, S. Gaspar, K. Hickman, K. Bishop, L. Milner, R. Krukar, and G. Petersen. "Scatterometry Applied to Microelectronic Processing." *Microlithogr. World* 1, no. 15 (1992): 16–22.
64. Raymond, C. "Scatterometry for Semiconductor Metrology." In *Handbook of Silicon Semiconductor Metrology*, edited by A. C. Diebold, 485–95. *New York*: *Marcel Dekker*, 2001.
65. Raymond, C. J., S. S. H. Naqvi, and J. R. McNeil, "Resist and Etched Line Profile Characterization Using Scatterometry." In *SPIE Microlithography Proceedings of SPIE*. Vol. 3050, 476–486, 1997; Jones, S. K., ed. "Metrology, Inspection, and Process Control for Microlithography XI." In *Proceedings of SPIE*. Vol. 3050, 476–486, July 1997.
66. Jekauc, I., J. Moffitt, S. Shakya, E. Donohue, P. Dasari, C. J. Raymond, and M. Littau. "Metal Etcher Qualification Using Angular Scatterometry." *Metrology, Inspection, and Process Control for Microlithography XIX*, edited by R. M. Silver, *Proceedings of SPIE*. Vol. 5752, July 2005.
67. Ukraintsev, V. A., M. Kulkarni, C. Baum, and K. Kirmse. "Spectral Scatterometry for 2D Trench Metrology of Low-*k* Dual-Damascene Interconnect." *Metrology, Inspection, and Process Control for Microlithography XVI*, edited by D. J. C. Herr, *Proceedings of SPIE*. Vol. 4689, 189–95, July 2002.

68. Bushman, S., and S. Farrer. "Scatterometry Measurements for Process Monitoring of Polysilicon Gate Etch." *Process, Equipment, and Materials Control in Integrated Circuit Manufacturing III*, edited by A. Ghanbari and A. J. Toprac. *Proceedings of SPIE*. Vol. 3213, 79–90, 1–2 October 1997 (August 1997).
69. Naqvi, S. S. H., R. H. Krukar, J. R. McNeil, J. E. Franke, T. M. Niemczyk, D. M. Haaland, R. A. Gottscho, and A. Kornblit. "Etch Depth Estimation of Large-Period Silicon Gratings with Multivariate Calibration of Rigorously Simulated Diffraction Profiles." *J. Opt. Soc. Am. A* 11, no. 9 (1994): 2485–93.
70. Naqvi, S. S. H., J. R. McNeil, R. H. Krukar, and K. P. Bishop. "Scatterometry and Simulation of Diffraction-Based Metrology." *Microlithogr. World* 2, no. 3, (1993).
71. Leray, P., S. Cheng, S. Kremer, M. Ercken, and I. Pollentier. "Optimization of Scatterometry Parameters for Shallow Trench Isolation (STI) Monitor." *Metrology, Inspection, and Process Control for Microlithography XVIII*, edited by R. M. Silver. *Proceedings of SPIE*. Vol. 5375, 2004.
72. Sendelbach, M., and C. Archie. "Scatterometry Measurement Precision and Accuracy below 70 nm." *Metrology, Inspection, and Process Control for Microlithography XVII*, edited by D. J. Herr, *Proceeding of SPIE*. Vol. 5038, 2003.
73. Smith, B., S. Bushman, and C. Baum. "Scatterometer Sensitivity for Statistical Process Control: Importance of Modeling for In-Direct Measurements." In *Characterization and Metrology for ULSI Technology 2005*, edited by D. G. Seiler et al., AIP Vol. 788, 437–41. 2005.
74. Adapted from input by Scott Bushman, Texas Instruments.
75. Baum, C., and S. Farrer. "Resist Line Width and Profile Measurement Using Scatterometry." *SEMATECH AEC-APC Conference*, Vail, CO, 1999.
76. Kostoulas, Y., C. J. Raymond, and M. Littau. "Scatterometry for Lithography Process Control and Characterization in IC Manufacturing." *Mater. Res. Soc. Symp. Proc.* 692, (2002).
77. Baum, C. C., R. Soper, S. Farrer, and J. Shohet. "Scatterometry for Post-etch Polysilicon Gate Metrology." *Metrology, Inspection, and Process Control for Microlithography XIII*, edited by B. Singh, *Proceeding of SPIE*. Vol. 3677, 148–58, 1999.
78. Chen, R. C.-J., F.-C. Chen, Y.-Y. Luo, B.-C. Perng, Y.-H. Chiu, and H.-J. Tao, "Application of Spectroscopic Ellipsometry Based Scatterometry for Ultra Thin Spacer Structure." *Metrology, Inspection, and Process Control for Microlithography XVIII*, edited by R. M. Silver, *Proceedings of SPIE*. Vol. 5375, 2004.
79. Raymond, C. J., M. Littau, B. Youn, C. -J. Sohn, J. A. Kim, and Y. S. Kang. "Applications of Angular Scatterometry for the Measurement of Multiply-Periodic Features." In *Metrology, Inspection, and Process Control for Microlithography XVII*, edited by D. J. Herr, *Proceedings of SPIE*, Vol. 5038, 2003.
80. Quintanilha, R., P. Thony, D. Henry, and J. Hazart. "3D-Features Analysis Using Spectroscopic Scatterometry." In *Metrology, Inspection, and Process Control for Microlithography XVIII*, edited by R. M. Silver, *Proceedings of SPIE*, Vol. 5375, 2004.
81. Reinig, P., R. Dost, M. Mort, T. Hingst, U. Mantz, J. Moffitt, S. Shakya, C. J. Raymond, and M. Littau. "Metrology of Deep Trench Etched Memory Structures Using 3D Scatterometry." In *Metrology, Inspection, and Process Control for Microlithography XIX*, edited by R. M. Silver, *Proceedings of SPIE*, Vol. 5752, 2005.
82. Rathsack, B. M., S. G. Bushman, F. G. Celii, S. F. Ayres, and R. Kris, "Inline Sidewall Angle Monitoring of Memory Capacitor Profiles." In *Metrology, Inspection, and Process Control for Microlithography XIX*, edited by R. M. Silver, *Proceedings of SPIE*, Vol. 5752, 2005.
83. Huang, H. T., and L. F. Terry Jr.. "Spectroscopic Ellipsometry and Reflectometry from Gratings (Scatterometry) for Critical Dimension Measurement and In Situ, Real-Time Process Monitoring." *Thin Solid Films* (2004): 455–6, see also 828–36.
84. Barna, G., L. M. Loewenstein, K. J. Brankner, S. W. Butler, P. K. Mozumder, J. A. Stefani, S. A. Henck., et al. "Sensor Integration into Plasma Etch Reactors of a Developmental Pilot Line." *J. Vac. Sci. Technol.* B12, no. 4 (1994): 2860.
85. Adapted from input by Charlie Kohn, SemiTest, Inc., Billerica, MA.
86. Epimet, SemiTest, Inc., Billerica, MA.
87. Brookside Software, http://www.brooksidesoftware.com

88. Barna, G. G. "Automatic Problem Detection and Documentation in a Plasma Etch Reactor." *IEEE Trans. Semicond. Manuf.* 5, no. 1 (1992): 56.
89. BBN Domain Corporation: Cambridge, MA.
90. Perception Technologies: Albany, CA.
91. Triant Technologies, Nanaimo, BC, Canada V9S 1G5.
92. Umetrics, Winchester, MA.
93. Brookside Software, San Mateo, CA.
94. Brooks Automation, Richamond, BC, Canada V7A 4V4.
95. Pattern Associates, Inc., Evanston, IL.
96. Real Time Performance, Sunnyvale, CA.
97. SEMATECH Fault Detection and Classification Workshop, February 18–19, 1997.
98. Barna, G. G. "Procedures for Implementing Sensor-Based Fault Detection and Classification (FDC) for Advanced Process Control (APC)." *Sematech Technical Transfer Document # 97013235A-XFR*, October, 10, 1997.
99. MiTex Solutions, Inc., Canton, MI.
100. ProcessWORKS, a member of the WORKS family of products being commercialized by Texas Instruments, Dallas, TX.
101. Barna, G. G. "APC in the Semiconductor Industry, History and Near Term Prognosis." In *SEMI/IEEE ASMC 96 Workshop*, 364–9. Cambridge, MA, November 14, 1996.

26

Yield Modeling

26.1 Introduction **26**-1
26.2 Cluster Analysis **26**-2
26.3 Yield Models **26**-4
Random-Defect Yield Models • General Yield Model
26.4 Yield Limits **26**-7
Random Defect Yield Limits • Systematic Yield Limits—Method 1 • Systematic Yield Limits—Method 2 • Test Yield Limits
26.5 Summary **26**-19
References **26**-19

Ron Ross
Texas Instruments, Inc.

Nick Atchison
Multigig, Inc.

26.1 Introduction

Yield modeling has been used for many years in the semiconductor industry. Yield models are now used not only for yield analysis, but also as the basis for automated yield analysis programs. Historically, the term "yield model" has referred to the mathematical representation of the effect of randomly distributed "defects" on the percentage of the integrated circuits (or die) on a wafer that are "good." Good means that they pass all parametric and functional tests that are specified for the product. The mathematical representations are typically derived from statistical distribution functions, such as the Poisson distribution or the Bose–Einstein statistics. Certain assumptions are then made about the variations in the spatial distributions of the "killing" defects on the wafers and mathematical formulas are derived from the results. References will be provided to yield analysis patents that use the yield models presented in this chapter.

The use of yield models that are based on statistical distributions is often successful for accurately calculating yields for products, which have their yield limited only by random defects. However, a complete and much more useful yield model will also account for "systematic" yield losses. Systematic yield losses can result from process, design or test problems. Die that fail for systematic problems are not randomly distributed over the wafer area, but are often confined to given regions, such as the outer periphery or the center. Systematic failures do not depend on die area, as do failures that are due to random defects.

The total yield for a given product can be expressed as the product of the systematic yield and the random yield:

$$Y = Y_s \times Y_r$$

Typically, the second term in this product is the only one that is "modeled," in the statistical model equation used to calculate the yield limits due to various types of random defects that arise from different manufacturing process steps or process equipment. Thus, the statistical modeling "partitions" the random yield limits (or, conversely, yield losses) into components that are due to different types

of defects. The term Y_s is left simply as a single factor and not partitioned. Y_s is often "estimated," or it can be calculated by performing "cluster analysis," which will be discussed in this chapter.

A complete yield model partitions the term Y_s into its sub-components to create a pareto of systematic losses so yield improvement projects can be prioritized. Two methods for doing this will be discussed in this chapter.

A complete yield model should, therefore, have the following characteristics:

1. It must account for all sources of yield loss, both random and systematic.
2. The total modeled or calculated yield should agree well with the actual yield.
3. It should ideally give insight into possible causes of the yield loss.
4. It should be able to partition and quantify yield losses resulting from design, process, test, and random defects.
5. It should provide the basic methods for automated yield analysis tools.

Yield modeling is worthwhile and advantageous because:

1. It makes possible the use of existing process and test data to quantify and paretoize all sources of yield loss.
2. It can substantially improve the yield learning rate for new products.
3. It makes accurate yield forecasting possible, which aids in planning.
4. It results in logical prioritizing of resources to work on yield enhancement projects with the highest payback.
5. It helps to set product specifications that match process capability.
6. It can provide the primary algorithms needed to create automated yield analysis programs. (See yield analysis patents in the References section.)

This chapter will first cover cluster analysis for calculating Y_s and Y_r. A brief review of the well-known random defect yield models will be given. The use of one of these models (the negative binomial model) for calculating individual defect yield limits will be explained. Two methods for calculating systematic yield limits will be discussed. The calculation of test yield limits will also be briefly presented.

26.2 Cluster Analysis

Cluster analysis or "window analysis," introduced by Seeds [1,2] was used extensively by Stapper at IBM [3]. This analysis is performed using actual wafer probe bin maps for finished wafers. The die are partitioned into groups or "blocks" of 1, 2, 3, 4 (2×2), 6 (3×2), 9 (3×3), etc. A simple example with groupings of 1×1, 1×2, 2×2, and 2×3 is shown in Figure 26.1. The percent yield is then calculated for each grouping, with the stipulation that the block is only considered to be a yielding block if all die within the block passed wafer probe testing (e.g., in the 2×2 block, all four dice must have yielded at wafer probe to count the block as good). For example, if there are 600 possible candidates on the wafer, and 480 tested good, the yield of the 1×1 block is simply 480/600 = 80%. For the 1×2 blocks, the total possible candidates would be 300, and if 216 of these had both die pass wafer probe, the yield is 216/300 = 72%. The 2×2 blocks, for example, have 150 total candidates, and if 90 of the blocks contain all four dice that tested good, the yield is 90/150 = 60%. This simulates what the yield would be for similar products of the same technology with 2×, 3×, 4×, 6×, and 9× the die area of the actual product being analyzed.

Because of statistical variations in yield across the wafer surface and from wafer to wafer, the above block yield calculations are performed on a relatively large number (~100–500 if possible) of wafers and the yields for each block size are averaged. When the averages have been computed, the values are plotted on a graph of yield vs. block size, as shown in Figure 26.2.

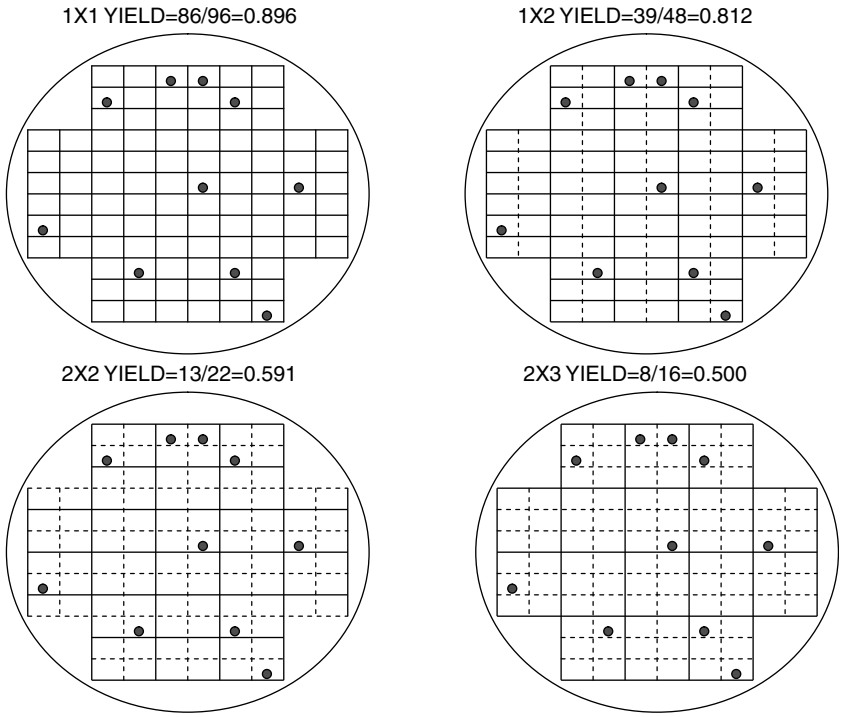


FIGURE 26.1 Cluster analysis groupings.

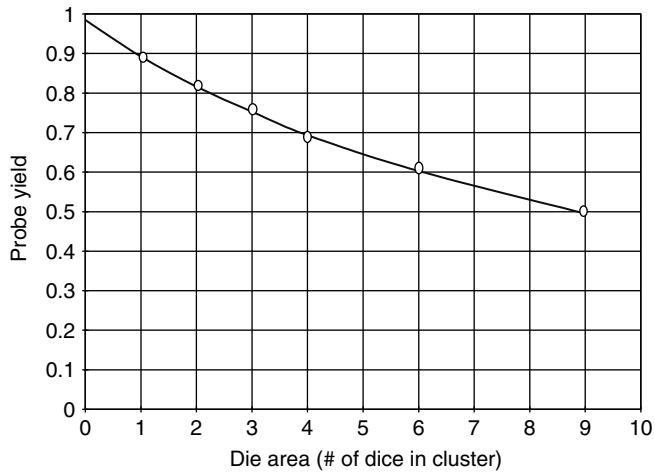


FIGURE 26.2 Cluster analysis graph.

A "best-fit" curve fitting routine is then applied to the points on the graph, using one of the yield model equations described later. For illustration, the negative binomial model will be used:

$$Y = Y_s \frac{1}{\left(1 + \frac{AD}{\alpha}\right)^{\alpha}} \tag{26.1}$$

where, Y_s (systematic yield limit); D (density of random fatal defects in units of defects per die); and α (cluster factor) are the parameters whose values are optimized for the best fit of the equation to the block yield data points.

The analysis thus provides "best" values for Y_s, D, and α (cluster factor) for a statistically significant sample of wafers for a given product. Here Y_s corresponds to the intercept of the best-fit curve to the y-axis and is the measure of the tendency of the defects to "cluster" or to depart from total randomness. D is the average number of fatal or killing defects per die. To calculate the killing defect density in terms of defects per unit area, D is divided by the die area of the product in question. The random defect yield limit (Y_r) can then be calculated either:

1. By using the equation

$$Y_r = \frac{1}{\left(1 + \frac{AD_0}{\alpha}\right)^{\alpha}} \tag{26.2}$$

   where D_0 is the defect density in terms of defects per unit area, or
2. By using

$$Y_r = \frac{Y}{Y_s} \tag{26.3}$$

   where Y is the average yield of the 1×1 block for the wafers used in the analysis.

If a statistically appropriate sample size was used for the analysis and if the negative binomial model provides a good fit, the two methods should agree very closely.

The use of cluster analysis provides two major benefits. The first is that it quantifies the systematic and random yield limits, so priority can be placed on further analysis and yield improvement efforts, which results in the greatest payback. The second is that it provides an excellent cross-check for the results of the yield modeling which partitions the random and systematic yield limits into their subcomponents. After the partitioning is completed, if all sources of yield loss are accounted for, the product of all random defect yield limits should equal the Y_r obtained from the cluster analysis. Likewise, the product of all independent systematic yield limits should equal the Y_s from the cluster analysis.

26.3 Yield Models

26.3.1 Random-Defect Yield Models

If it is assumed that a wafer has a given number of fatal defects that are spread randomly over the wafer area, then the average number per chip would be $A \times D_0$, where A is the chip area and D_0 is the total number divided by the total wafer area. If the defects are completely random in their spatial distribution, the probability of finding a given number k, of defects on any single die is given by the Poisson distribution:

$$P(k) = \frac{(\lambda^k \times \exp(-\lambda))}{k!} \tag{26.4}$$

where

$$\lambda = A \times D_0$$

The yield is then defined as the probability of a die having zero defects ($k=0$), so:

$$Y_r = P(0) = e^{-\lambda} = e^{-AD_0} \tag{26.5}$$

This is the Poisson yield model. In most cases, this model is found to predict yields that are lower than actual yields for a given product with a specified D_0. This is because the defect density varies by region of the wafer and from wafer to wafer. This results in a higher probability of a given die having multiple killing defects than would be predicted by the random model, thus leaving other dice without any killing defects which, in turn, means higher yields than predicted.

To take into account the variation in the defect density, the yield model is modified to include a defect density distribution:

$$Y_r = \int_0^\infty F(D)e^{-AD}dD \tag{26.6}$$

Several of the defect yield models used by wafer manufacturers result from solving this equation with various assumptions for the form of $F(D)$.

Of course, if $F(D)$ is a delta function at D_0, the Poisson model results.

If a Gaussian distribution is approximated by a triangular distribution as shown in Figure 26.3, the Murphy Model is obtained:

$$Y_r = \left(\frac{1-e^{-AD_0}}{AD_0}\right)^2 \tag{26.7}$$

An exponential distribution:

$$F(D) = \frac{1}{D_0}e^{-D/D_0} \tag{26.8}$$

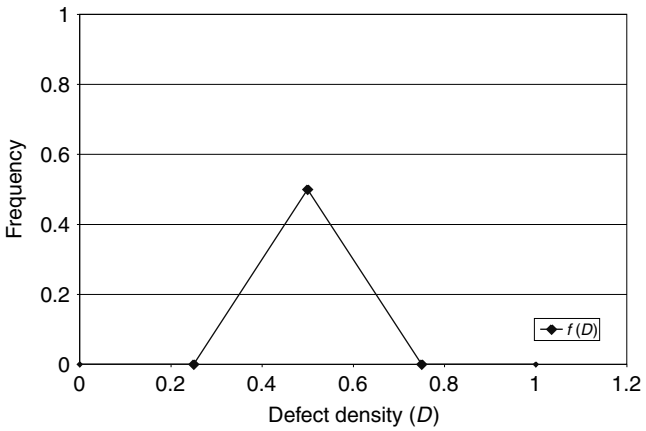


FIGURE 26.3 Triangular defect-density distribution.

results in the Seeds model:

$$Y_r = e^{-\sqrt{AD_0}} \tag{26.9}$$

Another common model is the Bose–Einstein model, which attempts to take into account the number of mask steps and assumes an average defect density of D_0 per mask level. This model takes the following form:

$$Y_r = \frac{1}{(1 + AD_0)^n} \tag{26.10}$$

where n is the number of mask levels. The main weakness with this model is that it assumes the same defect density for all levels. This is usually not the case. For example, the killing defect density is usually much higher for metal layers than for front-end layers. The Price model is a special case of the Bose–Einstein model which sets $n=1$.

The last model that will be considered is the negative binomial model. This has been discussed extensively by Stapper, formerly of IBM [4]. This can be derived by setting $F(D)$ in Equation 26.6 to the gamma distribution. The resulting random yield model is:

$$Y_r = \frac{1}{\left(1 + \frac{\lambda}{\alpha}\right)^{\alpha}} \tag{26.11}$$

where α is the cluster factor. This factor has physical significance and is related to the amount of clustering or non-randomness of the killing defects. A small value of α indicates a high degree of clustering and a larger value indicates a lower degree of clustering. A small value of α means that the probability of having multiple killing defects on the same die or on adjacent dice is greater than for a large α. Small α also means that the variation in D_0 is greater across the wafer area. As the value of α nears infinity, it can be shown that Equation 26.11 reduces to Equation 26.5. The determination of the value of α was explained in the previous section.

A comparison of the defect-yield limit predictions for the various models presented here is shown in Figure 26.4. It is seen that the Seeds model is the most pessimistic and the Bose–Einstein model is the

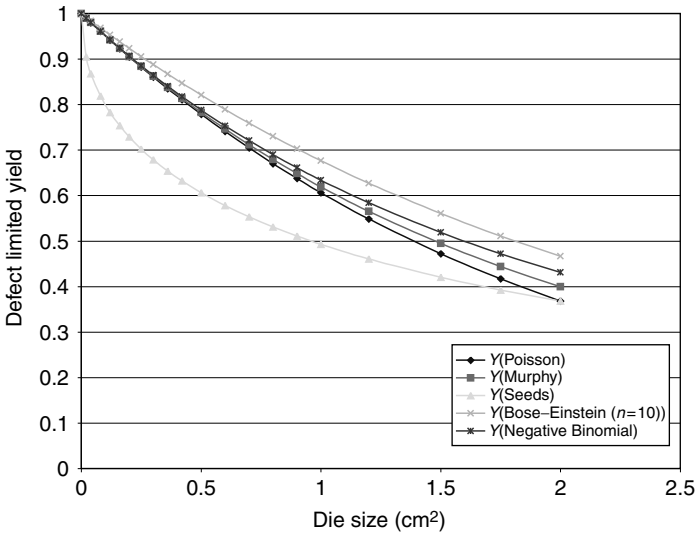


FIGURE 26.4 Random defect-limited yield vs. die size.

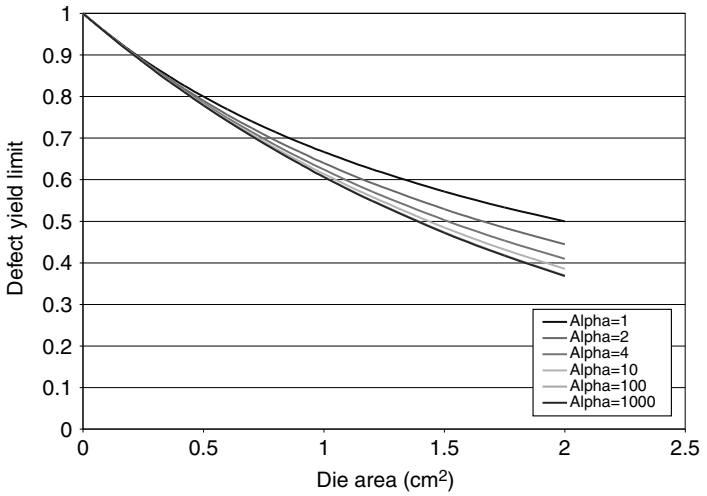


FIGURE 26.5 Negative binomial model—alpha comparison.

most optimistic. For these curves, a D_0 of 0.25/cm^2 has been assumed, and for the Bose–Einstien model, $n=8$ was assumed (so $D_0=0.25/8=0.03125$/cm^2/layer).

Figure 26.5 depicts Y_r curves for the negative binomial model for various values of α, again for $D_0=0.25$/cm^2.

It is seen from this graph that higher values of α result in a more pessimistic yield prediction.

26.3.2 General Yield Model

A general yield model can be expressed as a product of all independent yield limits

$$Y = \prod_i Y_{r_i} \prod_j Y_{P_j} \prod_k Y_{D_k} \prod_l Y_{T_l} \tag{26.12}$$

where

Y_{r_i} the random defect yield limit for defect type i;
Y_{P_j} the systematic process yield limit for process parameter j;
Y_{D_k} the systematic design yield limit for design sensitivity k;
Y_{T_l} the systematic or random yield limit for test problem l.

It is important to note that these yield limits must be independent of each other. This means that the various yield detractors do not affect each other or correlate with each other. If they do, double counting of yield problems ensues resulting in total yield calculations that are lower than actual yield. If some of the yield limits are interdependent, all but one of each interdependent group must be eliminated, or they must be proportioned according to the degree of correlation [5].

In subsequent sections, it will be explained how to calculate each of the various types of yield limits.

26.4 Yield Limits

26.4.1 Random Defect Yield Limits

The most accurate method for calculating yield limits resulting from random defects is to implement a thorough in-line inspection program using such tools as the KLA23xx.

There are two other requirements for obtaining accurate results:

1. Defect wafer map to wafer probe map overlay capability must exist, so it can be accurately determined upon which die each defect falls.
2. It is highly desirable to have a computer program that calculates the "kill ratio" or "killer probability" for each type of defect at each critical layer.

It is also important to have a consistent defect classification scheme, preferably automatic defect classification (ADC). This method works best for products in full production because fairly large numbers of wafers must be inspected ($\geq$100 wafers) to obtain accurate killer probabilities.

In order to get accurate and consistent results for these calculations, a large enough sample of the defects detected by the inspection equipment must be reviewed or classified. This sample size must be adjusted to take into account such considerations as inspection capacity and average numbers of defects on typical wafers. If the number of defects at a particular layer is small ($<\sim$100), then it is a good practice to review all of the defects. If the number is large, then at least 50–100 defects per wafer, per level, should be classified. It is very important that a random sample of the defects be classified, and not just the "large" ones or the "interesting" ones.

The killer probability is easily calculated as:

$$P_{k_i} = 1 - \frac{Y_d}{Y_c} \tag{26.13}$$

where

Y_d the probe yield of all of the die that had defects of type i at a particular layer, and
Y_c the probe yield of all of the die that did not have defects of type i.

Assume, for the sake of illustration, that 100 wafers were inspected at a given layer (e.g., poly), and that a total of 1000 dice were found with defects of a particular type (e.g., notching). Assume also that 9000 dice do not have defects of this type. Also, assume that, of the 1000 dice with defects, 300 of them were later tested to be good at wafer probe, and of the 9000 without defects, 6300 of them were later tested to be good at wafer probe. Now, the yield of the dice with defects (Y_d) would be 300/1000 = 30%, and the yield of the dice without defects (Y_c) would be 6300/9000 = 70%. The killer probability would then be:

$$P_{k_i} = 1 - \left(\frac{0.3}{0.7}\right) = 0.571 \tag{26.14}$$

After the killer probabilities have been calculated for all types of defects, the yield limits for each type of defect can be calculated by using the following equation:

$$Y_i = \frac{1}{\left(1 + \frac{P_{k_i} \times A \times D_i}{\alpha}\right)^{\alpha}} \tag{26.15}$$

where A, die area; D_i, defect density for defect type i; and D_i, is calculated from:

$$D_i = \frac{N_i}{A_T}$$

where, N_i, the total number of dice with defects of type i and A_T, total area inspected. There will, of course, be one of these yield limit equations for each different type of classified defect at each inspected layer.

It is important to note that N_i is the "normalized" number of defects of type i. This means that, if fewer than 100% of the defects are classified (or reviewed), the total number of defects of type i is projected

onto the total population by the equation:

$$N_i = \frac{N_{i_R}}{N_R} N \tag{26.16}$$

where

N_{i_R} the number of defects of type i in the reviewed sample
N_R the total number of defects reviewed (of all types); and
N the total number of defects detected (of all types).

It must also be noted that only Equation 26.15 works for unclustered (or random) defects.

26.4.2 Systematic Yield Limits—Method 1

The first method for calculating individual systematic yield limits is called Limited Yield Analysis. This method only gives accurate results for products in high volume production. This is because there are many variables that affect yield, and this method uses wafer averages, and determines the average effect of electrical parameters (e.g., threshold voltages or transistor gains) on the yield. The actual number of wafers required to calculate a yield limit in the range of 0.97 (3% loss) and with a standard deviation of about 8% in the yield of the product has been found empirically to be about 2000 wafers. For higher yield limits (less loss than 3%), or for a higher standard deviation, the number must be even greater.

The method works for either process or design-related systematic problems.

The yield limits are calculated from the formula:

$$Y_{D_k} = \int_{P_{k\mathrm{MIN}}}^{P_{k\mathrm{MAX}}} F(P_k) Y(P_k) \mathrm{d}P_k \tag{26.17}$$

where

Y_{D_k} design yield limit due to electrical parameter k;
P_k values of electrical parameter k;
$F(P_k)$ normalized distribution function of parameter k; and
$Y(P_k)$ normalized probe yield as a function of parameter k.

Equation 26.17 also works for Y_{P_k} (process systematic yield limits).

A detailed explanation of how this integral is evaluated is in order. Using specific examples is the best way to do this.

Figure 26.6 shows a histogram (or frequency distribution) for a particular electrical parameter (poly sheet resistance in this example) for a product manufactured using a BiCMOS process. The specification limits for this parameter are 800–1200 ohms/square. The data results from measuring poly sheet on five sites per wafer on a large number of wafers. The histogram represents the average values of the five sites for each wafer. To use this distribution in Equation 26.17, it must be normalized so that the total area of the histogram is 1.0. This is done by simply dividing the number of wafers in each range (or bar) by the total number of wafers in the sample.

The key element in the analysis is now the "grouping" of the wafers into three groups with equal numbers of wafers in each group. The first group consists of all wafers with average poly sheet in the lower third of the distribution. The second group consists of wafers with medium values of poly sheet. The third group includes all wafers with high values of poly sheet. Again, the groups are not formed by equal ranges of poly sheet, but by equal numbers of wafers in each group. After the wafers are grouped in this manner, the average of the electrical parameter (poly sheet) is computed for each group. Also, the average of the wafer probe yield is computed for each group. This results in three points that can then be plotted as shown in Figure 26.7.

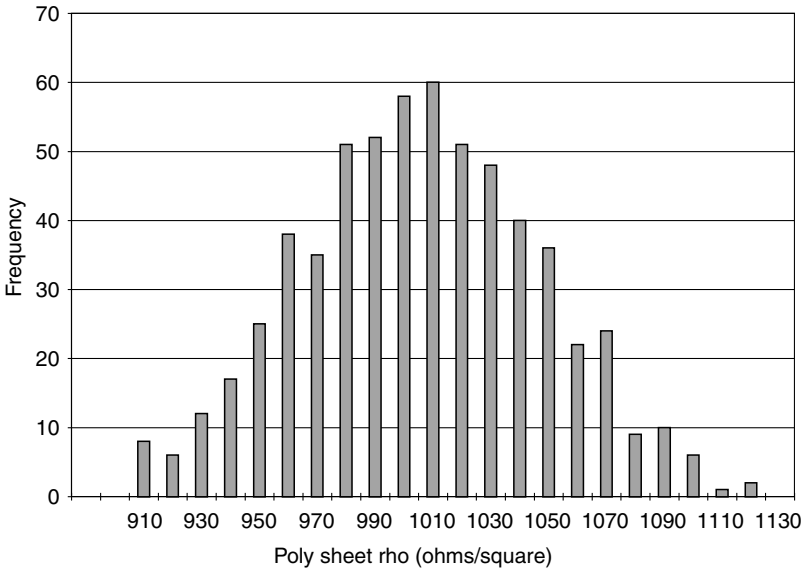


FIGURE 26.6 Poly sheet resistance frequency distribution.

A function (in this case, a parabola) can then be fitted to these three points. The function must then be normalized so that the highest yield of the three points is equal to 1.0. The normalized values for the other two points are then computed by dividing the average number of good dice for the two groups by the average number of good dice for the highest yielding group. For example, if the highest average yield is 400 DPW, and the other two averages are 360 and 370 DPW, the corresponding normalized values

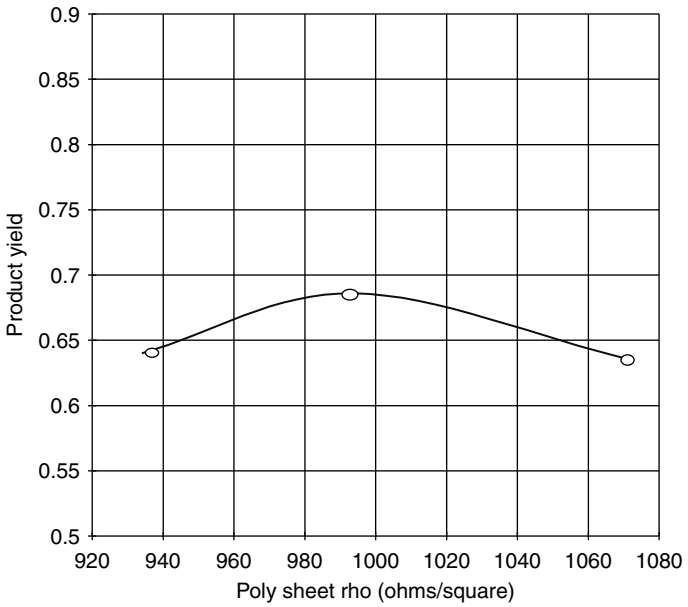


FIGURE 26.7 Mean product yield vs. poly sheet resistance.

would be 360/400 = 0.90 and 370/400 = 0.925. After the curve has been normalized in this manner, the resulting function becomes $Y(P_k)$ in Equation 26.17. Equation 26.17 can then be evaluated (numerically because of the irregularity of the frequency distribution) to compute the corresponding yield limit. In this special example, the yield limit was calculated to be 0.936, which translates into a yield loss of 6.4% because of the probe yield sensitivity to poly sheet.

It is important to understand why the electrical test data and the yield data are grouped and averaged as explained above. There are hundreds of variables that might affect wafer probe yields, including both random and systematic variables. As long as these variables are independent of the variable in question (poly sheet in the example), there is no reason to expect that the other variables should group themselves in any particular way within the three groups partitioned according to the value of poly sheet. In other words, there is no reason to expect that wafers with low poly sheet should have more defects of a particular type than wafers with high poly sheet, for example. Likewise, for example there should be no reason to expect that wafers with low poly sheet would have a different value of contact resistance compared to wafers with high poly sheet. Therefore, if the sample size is large enough, all of the other independent variables should average out in a similar manner for each of the three groups of wafers, leaving only the effect on yield of the variable being analyzed. Essentially, the averaging is a method for looking at the variation in yield due to one parameter, where all other effects being equal for the three groups, have been "removed" from the analysis. It is simply cleaner than trying to look at and interpret a scatter plot of the raw data.

If two electrical parameters correlate with each other (e.g., NMOS effective channel length (L_{eff}) and NMOS drive current), then when the wafers are grouped according to ranges of one variable, the other parameter will also show a systematic grouping. For example, the wafers with low L_{eff} would also tend to have higher drive current. If the limited yield analysis shows that one parameter affects yield, the other correlating parameter will also affect yield. If the correlation is very good, the yield limits should be very nearly equal for the two parameters. In generating the yield limit pareto, only one of the two yield limits would be used. The one to be used must be chosen by engineering judgement, based on knowledge of the product in question.

Regarding the use of just three groupings of the data, two justifications are in order. Three points are the minimum required for determining the shape of the curve, whether it is linear or has a peak and drops off on both sides (as in Figure 26.7). Normally, a more complex relationship (e.g., two peaks) would not be expected. The second reason for using three points is because this maximizes the number of wafers in each group (compared to using more than three points) so the averaging described above is most effective for detecting significant yield differences among the groups. In other words, a 3% yield difference, for example, would be more significant with three groups than with more groups because more data points are included in each group.

In general, if the yield varies by value of an electrical parameter, such as transistor parameters, resistance or capacitor values while the said parameters are within their specification limits, a design issue is indicated. If the yield varies by electrical test values of leakage current between metal or poly lines, or current leakage between nodes of transistors (e.g., emitter-to-base leakage, etc.), a process problem is indicated.

26.4.3 Systematic Yield Limits—Method 2

In this section, a powerful method for determining systematic yield limits called product sensitivity analysis (PSA) is presented. This has been described in detail by Ross and Atchison [6]. This has three advantages over the previous method. The first is that significantly fewer wafers are required for the analysis. This is because the analysis uses site data (by x–y location on the wafers) as opposed to wafer averages. The second advantage is that the analysis can be done earlier in the development cycle, after only ~100 wafers have been processed and tested. This makes it possible to detect systematic process and design problems early and fix them before the product is shipped to customers. The third advantage is

that PSA gives insight into possible causes of the yield loss. This will be explained in more detail as the analysis is described.

Product sensitivity analysis determines how parameters measured at wafer probe (e.g., I_{cc} currents, offset voltages, cutoff frequencies, etc.) vary with electrical parameters (e.g., V_{tn}, L_{eff}, H_{fe}, poly sheet, etc.). This method does not work for functional pass/fail tests performed at wafer probe.

In preparation for the analysis, a wafer-probe test-yield loss pareto is generated as shown in the example of Figure 26.8. The tests for which actual parametric values are measured and which cause the greatest yield loss are then chosen for the analysis. The analysis can then be performed for these parametric tests vs. all electrical parameters. The example used here will be for the wafer probe test highest on the pareto (test 2304). Figure 26.9 shows the actual distribution of this parameter. It can be seen that the distribution is far off-center between the specification limits. It is required that all parameters measured at wafer probe be recorded and stored by *x–y* coordinates on the wafer. Also, electrical parameters must be measured either in scribe-line test modules or in drop-in test modules at several sites on the wafers (preferably at least five per wafer).

The analysis then proceeds as follows for each wafer probe parameter. The wafer probe and electrical parameters are screened and data points outside the "normal" or "typical" distributions are excluded. This may be done, for example, using:

$$U_{SL} = Q3 + \frac{3}{2} \times \text{IQR} \tag{26.18}$$

where

$$\text{IQR} = Q3 - Q1$$

and

L_{SL} low screen limit, and
U_{SL} high screen limit

Next, the values of the probe parameter for the product dice immediately surrounding each electrical test site are averaged and paired with the corresponding values for all electrical parameters of the sites. This results in a table similar to the abbreviated hypothetical example shown in Table 26.1. The first column identifies lot, wafer number and site. The next column is the average value for the wafer probe parameter for dice surrounding each electrical test site. The following columns contain the values for the electrical parameters.

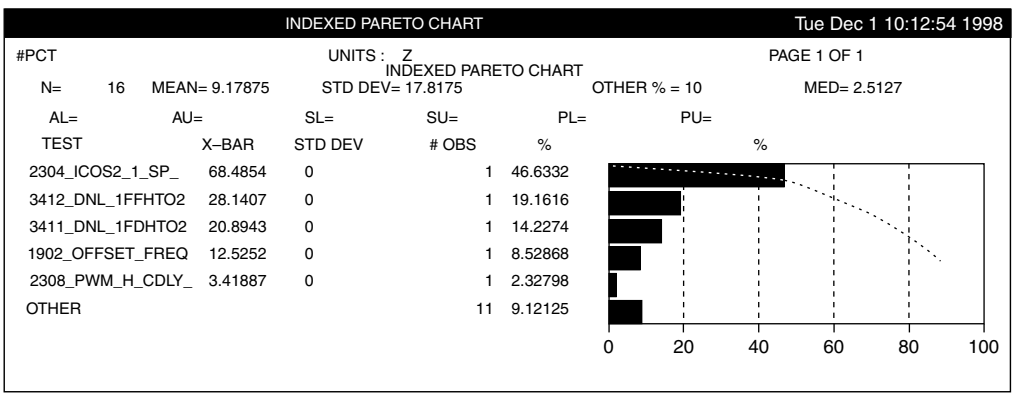


FIGURE 26.8 Wafer-probe test-yield loss pareto analysis.

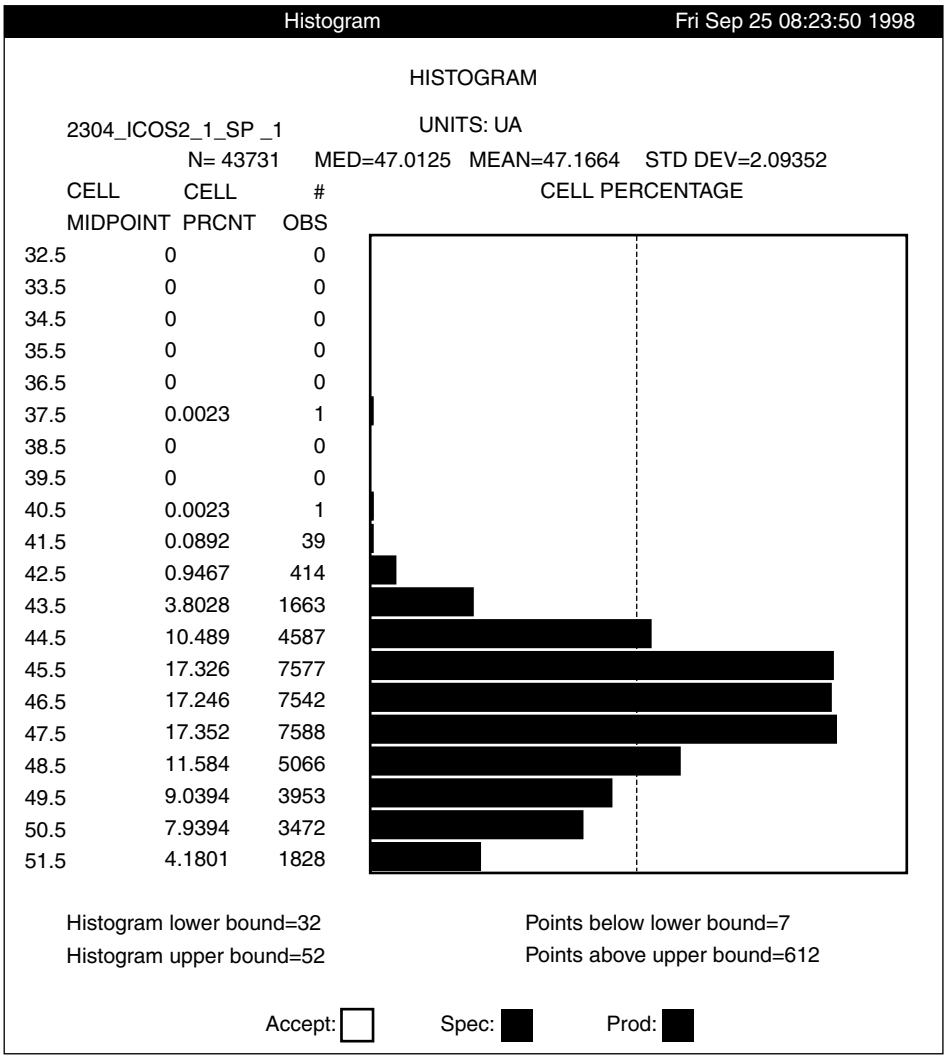


FIGURE 26.9 Frequency distribution for test 2304 (highest contributor to yield loss in Figure 26.8 pareto).

For each electrical parameter, the data are then grouped into three ranges, each with equal numbers of sites regardless of which wafer or lot they come from, according to the value of the electrical parameter. For the three groups, the average of the electrical parameter is computed, as is the average of the wafer probe parameter. This gives three points to plot on a graph, just as in the case of the limited yield analysis. A best-fit line is then calculated and plotted on the graph.

The upper and lower specification limits for both the electrical parameter and the wafer probe parameter must also be imported into the computer program. These are plotted as vertical lines for the electrical parameter (independent variable) and horizontal lines for the wafer probe parameter (dependent variable).

The standard deviation (σ) is computed for the wafer probe parameter for each of the three groups. The $\pm 3\sigma$ bars are plotted for each of the three points on the graph, and the $\pm 3\sigma$ lines are plotted above and below the best-fit line. These are forced to be parallel to the best-fit line. The result is a graph like

TABLE 26.1 Parametric and Product Dice Date Table

Lot_Wafer_Site	Mean of Probe Test 2034	Poly Sheet	NMOS L_{eff}	NPN Beta
L107_1_1	78	194	0.65	85
L107_1_2	74	201	0.68	92
L107_1_3	73	198	0.64	83
L107_1_4	73	202	0.63	88
L107_1_5	73	203	0.64	86
L107_2_1	74	196	0.66	88
L107_2_2	72	201	0.68	98
L107_2_3	71	199	0.63	85
L107_2_4	73	201	0.64	91
L107_2_5	72	203	0.65	93
L107_3_1	73	197	0.63	82
L107_3_2	75	200	0.67	90
L107_3_3	74	197	0.62	84
L107_3_4	71	203	0.65	85
L107_3_5	70	205	0.63	84
L108_1_1	80	190	0.61	88
L108_1_2	79	192	0.65	96
L108_1_3	81	190	0.62	85
L108_1_4	80	191	0.63	86
L108_1_5	77	193	0.62	87
L108_2_1	78	194	0.63	86
L108_2_2	72	199	0.67	93
L108_2_3	80	192	0.64	85
L108_2_4	73	197	0.64	89
L108_2_5	73	198	0.62	90

the example shown in Figure 26.10. This is for probe test 2304, whose distribution is shown in Figure 26.9. Note that the frequency distributions for the electrical parameter are also plotted on the graph.

If the slope of the best-fit line is large relative to the wafer probe parameter spec limits, a sensitivity is indicated. Of course, a statistical test (such as *t*-test) should be performed to confirm that the differences in the wafer probe parameter among the three groups are statistically significant at the 95% level. Also, the three averages for the wafer probe parameter should monotonically increase or decrease with increasing value of the electrical parameter. If these conditions are met, the sensitivity is of interest and should be further studied.

If the $+3\sigma$ line intersects the upper spec limit of the wafer probe parameter while the electrical parameter is still within its spec limits, yield loss can result for sites with electrical parameter values between this intersection and the specification limit. Yield loss can similarly occur if the -3σ line intersects the lower specification limit of the wafer probe parameter at a point between the electrical specification limits. In Figure 26.10, yield loss would start to occur near sites with values of poly2 sheet resistance (P2MSRES) below about 188 ohms/square. This value is well within the specification limits of 160–240 ohms/square.

For electrical parameters such as V_{tn}, H_{fe}, poly sheet, etc., where specification limits were set in advance and agreed upon by the designers, if the line intersect as described in Figure 26.10, any yield loss is due to design sensitivities. Of course, if the electrical parameter is tightly controlled in a narrow range where the $+3\sigma$ lines are within the probe specification limits, no yield loss would occur even though there is a design sensitivity.

The method for calculating yield limits is illustrated in Figure 26.11. Here, the best-fit line intersects the upper spec limit of the wafer probe parameter within the actual distribution range of the electrical parameter. A vertical line is extended to the horizontal axis. This line cuts off a certain fraction of the distribution. The yield loss is then simply the ratio of the area of the distribution to the left of the line

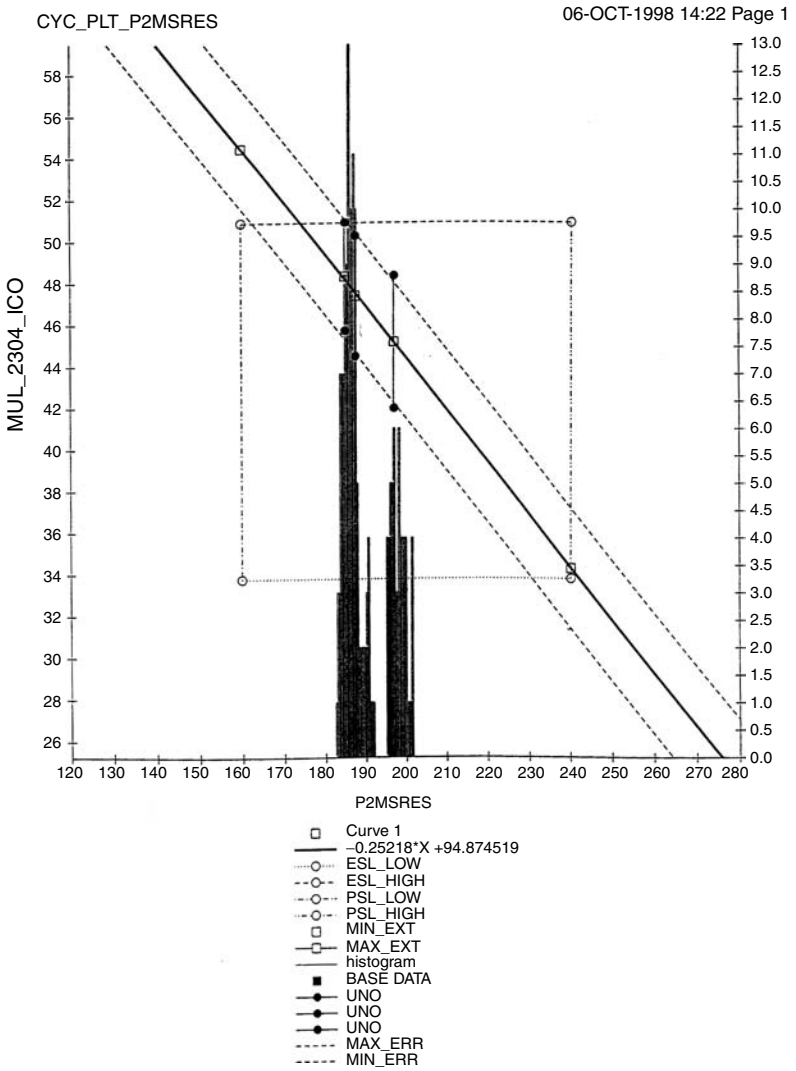


FIGURE 26.10 Best-line fit (and $\pm 3\sigma$) for "3-point analysis" of test 2034.

(in this example) to the total area of the distribution. The yield limit is then:

$$Y_{D_k} = 1 - \frac{A(F_{X_l})}{A(F_{X_t})}$$

where

Y_{D_k} design yield limit due to electrical parameter k;
$A(F_{X_l})$ the area under the distribution to the left of the vertical line; and
$A(F_{X_t})$ the total area under the distribution

This analysis is repeated for all electrical parameters and for each of the wafer probe parameters high on the yield-loss pareto. All of the graphs with no intersection of the $\pm 3\sigma$ lines with the wafer probe specification limits can then be eliminated. Also, graphs that do show potential yield loss because

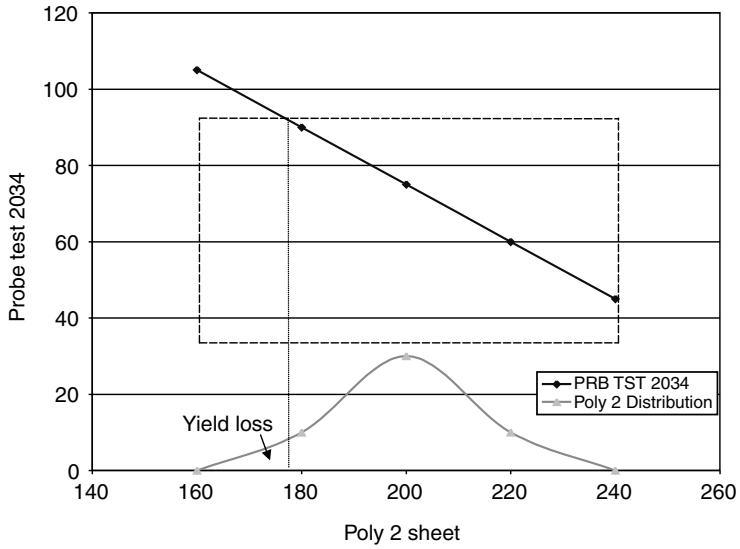


FIGURE 26.11 Product sensitivity analysis graph: probe test 2034 vs. poly2 sheet resistance—yield limit calculation.

of sensitivities to electrical parameters that correlate with other electrical parameters are screened, as described for limited yield analysis. This leaves only one yield limit per sensitivity that is independent of all others. The systematic yield limits are then summarized in a pareto.

26.4.4 Test Yield Limits

Test yield loss can result from such problems as mis-calibration of test hardware, poor contact of probe tips, electrical noise, etc. The calculation of two types of yield limits will be described here.

The first type is typically systematic in nature and results from differences between various testers, probe cards, interface boards, etc.

A "set-up" is defined as a unique combination of tester, probe card, interface board, and any other component that might be changed periodically from one tester to another.

The number of wafers that should be run using each set-up in order to obtain statistically valid results depends on the variability in yield and the magnitude of the yield difference to be detected. One formula that can be used if the yield distribution is reasonably close to normal is:

$$N = (Z_\alpha + Z_\beta)^2 \frac{\sigma^2}{\delta^2}$$

where

Z_α Z factor for the probability of α of making an α type error;
Z_β Z factor for the probability of β of making a β type error;
σ standard deviation in the yield; and
δ the yield difference to be detected.

The z factors can be looked up in a standard table for the normal distribution. For example, if it is desired to have less than 10% probability of making an error of either type (projecting a yield difference when there is none, or projecting no yield difference when there is one), z is read from the table as 1.64. This assumes that the yield difference can be either positive or negative. If the σ for the yield is 8% and it

is desired to detect a yield deference of 1%, N can be calculated as:

$$N = (1.64 + 1.64)^2 \frac{0.08^2}{0.01^2} = 689 \text{ wafers}$$

As an example, it is assumed that three different set-ups have been used to test 3000 wafers with the following results:

Set-Up	N	% Yield
1	1200	87
2	800	85
3	1000	88

It is assumed that set-up "3" gives the optimum yield for this product. The test set-up yield limit is then:

$$Y_{T_S} = \frac{1200 \times 0.87 + 800 \times 0.85 + 1000 \times 0.88}{3000 \times 0.88} = 0.986$$

The general formula can be expressed as:

$$Y_{T_S} = \frac{\sum N_i \times Y_i}{N \times Y_{OPT}}$$

where

N_i the number of wafers tested with set-up i;
Y_i mean yield of wafers tested with set-up i;
N total number of wafers tested; and
Y_{OPT} optimum yield (or mean yield of the best set-up.

The number of wafers required for this type of yield calculation can be reduced substantially by splitting lots at wafer probe and testing the splits on different set-ups. With an added test of statistical significance (e.g., the F test), a yield difference could then be detected with only one lot.

Another type of yield loss can happen because of some product dice having one or more parameters measured at wafer probe that are close to the upper or lower specification limits. In this case, electrical noise or slight variations in tester voltage or current input levels can cause the dice to fail when they are really good. If the dice are tested repeatedly, they sometimes pass and sometimes fail. These dice are sometimes called "twinkling die." If correlation wafers are tested repeatedly to ensure proper test set-ups, the probe results can be used to calculate yield limits due to tester non-repeatability or twinkling dice. The formula is:

$$Y_{T_T} = 1 - \frac{\sum n_i T_i}{N_G \times N_t}$$

where

n_i the number of twinkling dice that fail i times;
T_i the number of times these dice failed;
N_G the gross number of dice in the (correlation) wafer; and
N_t the number of times the wafer was tested

For a simple example, it is assumed that $N_G = 600$ and that the correlation wafer was tested four times ($N_t = 4$). Furthermore, the number of dice that failed various numbers of times are assumed to be:

T_i	n_i
1	27
2	9
3	1

therefore,

$$Y_{T_T} = 1 - \frac{1 \times 3 + 9 \times 2 + 27 \times 1}{600 \times 4} = 0.980$$

The calculation of a few types of yield limits has not been described in detail in this chapter. For example, special yield test structures that are run early in the technology development cycle can be used to determine defect densities at various layers. This information can be used, along with critical area analysis, as soon as the first product layout is complete, to calculate defect yield limits for the new product. Also, yield limits can be calculated for clustered defects, but these must be treated separately from the defects that are random [7].

TABLE 26.2 BCA IYM Activities Summary Sheet

	Actual Q1CY97	PRML March	PRML April	PRML May	Active Projects
Design/product yield					
Yb (Beta)	0.994	0.995	0.992	0.998	DOE, characterization and spec adjustment
Y1 (L_{eff})	0.994	0.992	0.996	0.998	DOE, characterization and spec adjustment
Yr (P1000/250)	0.991	0.991	0.997	0.999	DOE, characterization and spec adjustment
Design/prod limiter	0.979	0.978	0.985	0.995	
Test limited yield Yt (test)	0.982	0.986	0.986	0.986	Probe card/setup improvements/OLI/test robustness
System process limiter					
Yi (IEBS)	0.995	0.995	0.992	0.996	
Ym1 (Metal 1)	0.993	0.985	0.998	0.997	
Ym2 (Metal 2)	0.996	0.997	0.994	0.996	
Ym (Metal)	0.989	0.982	0.992	0.993	
Yv(Via)	0.997	1	1	0.997	
Yc (Contact)	0.997	1	0.991	0.999	
System procs limiter	0.978	0.977	0.975	0.985	
Random defect limiter					SRAM strip back analysis
M2 bridge particle	0.986	0.985	0.984	0.985	
M2 bridge no part	0.996	0.996	0.995	0.996	
M2 extra pattern	0.999	0.999	1	1	
Metal 2 total	0.98	0.98	0.979	0.981	IDIRT2 defect reduction team
M1 bridge particle	0.985	0.984	0.984	0.9812	
M1 bridge no part	0.994	0.9948	0.9924	0.9904	
M1 extra pattern	0.995	0.996	0.999	1	
Metal 1 total	0.973	0.975	0.976	0.972	Metal 1 defect reduction team
Metal 0	0.994	0.996	0.997	0.992	
Scratches	0.974	0.973	0.978	0.98	Mechanical damage and scratch reduction
Residual metal	0.999	0.999	1	1	
Poly bridging	0.997	0.998	0.997	0.997	
Missing Co/Via	0.999	0.999	0.999	1	
Random defect limiter	0.919	0.922	0.927	0.924	
Calculated yield	0.863	0.869	0.878	0.893	
Actual yield	0.861	0.889	0.87	0.889	

26.5 Summary

A complete mathematical yield model has been discussed which accounts for yield limits resulting from design, process, test and random defect problems. The yield model is expressed as the product of all of the independent yield limits. Table 26.2 gives an actual example of all of the calculated yield limits for a product that was in production several years ago. The grand limit is shown and compared with the actual yield for a given month. It can be seen that the numbers agree very well. This means that nearly all of the sources of yield loss were accounted for.

Equations and methods have been described for calculating yield limits for most of the types of yield loss that occur in the four main categories. The effectiveness of the methodology has been shown by the example in Table 26.2.

References

1. Seeds, R. B. "Yield, Economic, and Logistic Models for Complex Digital Arrays." *IEEE International Convention Record* Part 6 (1967): 61–6.
2. Seeds, R. B. "Yield and Cost Analysis of Bipolar LSI," Presented at the 1967 International Electron Device Meeting Keynote Session, October 1967.
3. Stapper, C. H. "Large-Error Fault Clusters and Fault Tolerance in VLSI Circuits: A Review." *Journal of Research and Development* 33, no. 2 (1989): 162–73.
4. Stapper, C. H. "Fact and Fiction in Yield Modeling." *Microelectronics Journal* 20, no. 1–2 (1989): 129–46.
5. Ross, R., and N. Atchison. "A Useful Method for Calculation of Parametric Yield Limits." *Texas Instruments Technical Journal* 15, no. 4 (1998): 74–7.
6. Ross, R., and N. Atchison. "A New Method for Analyzing Probe Yield Sensitivities to IC Design." *Texas Instruments Technical Journal* 15, no. 4 (1998): 78–82.
7. Ross, R., and N. Atchison. "The Calculation of Wafer Probe Yield Limits from In-Line Defect Monitor Data." *Texas Instruments Technical Journal* 15, no. 4 (1998): 83–7.
8. Atchison, N., and R. Ross. Patent granted U.S. PTO No. 6,393,602, May 21, 2002.
9. Atchison, N., and R. Ross. Patent granted UP PTO No. 6,324,481, Nov. 27, 2001.
10. Atchison, N., and R. Ross. Patent granted U.S. PTO No. 6,210,983, Apr. 3, 2001.
11. Atchison, N. Yield Analysis Web Page holomirage.com

27

Yield Management

27.1 Introduction **27**-1

27.2 Sources and Types of Random Defects **27**-2
Diffusion and Implant • Surface Preparation • Photolithography • Etch Deposition and Oxide or W Chemical–Mechanical Polishing (CMP) • Cu CMP and Damascene Multi-Level Metalization • Wafer Edge Engineering

27.3 Yield Management Methodology **27**-8
Management Priority and Motivation • Process and Equipment Control • Product-Based Defect Detection and Analysis (Line Monitor) • Yield Impact Prediction/Verification • Root Cause Isolation

27.4 Summary **27**-28

References **27**-28

Louis Breaux
Sean Collins
Texas Instruments, Inc.

27.1 Introduction

The remarkable progress in integrated circuit manufacturing can, in no small part, be attributed to the evolution and advancement in the area of yield enhancement and yield management in the past decades [11]. This progress has been recognized not only just within the wafer manufacturing community but also by the wafer packaging community [12]. Considering imitation is the greatest compliment, then the fact that the yield management lessons and methodologies from the manufacturing "front end" are being adapted and applied to "back end" processes such as wafer bump and packaging is indeed an acknowledgement of the tremendous successes achieved [12].

Contamination-free manufacturing (CFM) originated as a term in the late 1980s. It was intended to describe the practice of semiconductor manufacturing under "ultraclean" conditions resulting in "perfect" yields. However, it is self-evident that perfect yields and ultraclean processes are goals that one strives for, but never achieves in totality under the constraints of time and money. Hence, the discussion in this chapter is primarily directed towards the methodology of "yield management" and the CFM practices that are relevant in achieving the highest yields possible in the shortest time.

Probe yield can be defined as:

$$Y = Y_s Y_r \tag{27.1}$$

In this definition, $(100 - Y_s)$ is the percentage of yield lost to systematic issues, which are not randomly distributed and tend to impact all or most die on a wafer. The process/design marginalities, parametric test conditions or reticle defects are common causes of systematic yield loss. They are typically encountered in the early phase of new device qualification and once addressed, tend not to recur. The percentage of yield lost to random defects on the wafer surface $(100 - Y_r)$ is mostly due to contamination.

TABLE 27.1 Chip Size Trends

Year of First Product Shipment	2003	2005	2007	2010	2013	2016	2018
Technology Generations Min Dimensions							
Dynamic random access memory (DRAM) half-pitch (nm)	100	80	65	45	32	22	18
MPU/ASIC M1 half-pitch (nm)	120	95	76	54	38	27	21
MPU printed gate length (nm)	65	45	35	25	18	13	10
Functions/Chip							
DRAM bits/chip-generation	1G	1G	2G	4G	8G	32G	32G
DRAM gbit/cm^2 at production	0.77	1.31	2.22	5.19	10.37	24.89	39.51
MPU functions per chip at introduction- million transistors (mtransistors)	180	285	453	1546	3092	6184	9816
MPU functions per chip at production- (mtransistors)	153	243	386	773	1546	3092	4908
High performance MPU functions per chip- (mtransistors)	439	697	1106	2212	4424	8848	14,045
Chip Size (mm^2)							
DRAM-at production (mm^2)	139	82	97	83	83	138	87
DRAM-at introduction (mm^2)	485	568	662	563	560	464	292
MPU-at production (mm^2)	140	140	140	140	140	140	140
MPU-at introduction (mm^2)	280	280	280	280	280	280	280
High Performance MPU at production (mm^2)	310	310	310	310	310	310	310
Lithographic field size (mm^2)	22×32 704	22×32 704	22×32 704	22×32 704	22×32 704	22×32 704	22×32 704
Maximum Substrate Diameter (nm)							
Bulk or epitaxial or SOI[a] wafer	300	300	300	300	450	450	450

Source: *The International Technology Roadmap for Semiconductors*, International SEMATECH, Austin, TX, 2003 Edition.
[a] SOI-silicon on insulator.

$$Y_r = e^{-AD} \tag{27.2}$$

where A is the area of the die and D is the defect density. Chapter 29, describes various methods used to calculate Y_s and Y_r for a given process flow and device. In modern factories using sub-0.25 μm design rules, 40%–50% of the total yield loss in the first year of production can be attributed to random defects. Prior chapters have dealt with process and design requirements for eliminating systematic yield losses. This chapter is primarily focused on the management of these random defect sources.

The international technology roadmap for semiconductors (ITRS) National Roadmap for Semiconductors shows the overall industry trends in design rule and chip size (Table 27.1). These trends have direct impact on yield as seen in Table 27.2 [1]. Shrinking design rules imply smaller killing defects, larger chip sizes cause lower yields for comparable defect densities, and more process steps means more sources of contamination/defects. The ITRS roadmap (Table 27.2) estimates significant and necessary reductions in unit process defect densities (microprocessor unit (MPU) Random particle per wafer pass) to achieve 75% MPU yield (when compared to a 90 nm design rule microprocessor at 75% yield) in the first production year of 22 nm design rule microprocessor.

27.2 Sources and Types of Random Defects

In the cast of bad actors, particles invariably play the lead role. Ionic, metallic, and organic impurities as well as trace amounts of moisture or oxygen also commonly cause random defects. Contamination can result from virtually all aspects of semiconductor manufacturing including processes, equipment, raw materials (chemicals, gases, wafers), fluid storage and delivery systems, wafer transport and storage (cassettes, stockers, and wafer boxes), cleanroom and people. Equipment, processes, and process materials contribute a majority of contamination to the wafer surface. Contamination from cleanroom,

TABLE 27.2 Yield Model and Defect Budget Technology Requirements

Year of first product shipment	2003	2005	2007	2010	2013	2016	2018
DRAM half-pitch (nm)	100	80	65	45	32	22	18
MPU/ASIC M1 half-pitch (nm)	120	95	76	54	38	27	21
MPU Printed gate length (nm)	65	45	35	25	18	13	10
Critical defect size (nm)	54	40	33	23	16	11	9
DRAM Random defect D_0 at production chip size and 89.5% (faults/m^2)	2216	3751	3190	3722	3722	2233	3545
MPU Random defect D_0 at production chip size and 83% (faults/m^2)	1395	1395	1395	1395	1395	1395	1395
Electrical D_0(faults/m^2) at critical defect size or greater 75% Yield+	2210	2210	2210	2210	2210	2210	2210
Chip size (mm^2)	140	140	140	140	140	140	140
# Mask levels - MPU	29	33	33	35	35	39	39
Random faults/mask	48	42	42	40	40	36	36
MPU Random particles per wafer pass (PWP) budget (defects/m^2) for generic tool type scaled to 54 nm critical defect size or greater							
CMP clean	397	195	129	58	29	12	8
CMP insulator	961	472	312	141	71	30	20
CMP metal	1086	534	352	159	81	34	23
Coat/develop/bake	174	85	56	25	13	5	4
CVD insulator	854	420	277	125	63	27	18
CVD oxide mask	1124	552	364	165	83	35	24
Dielectric track	273	134	89	40	20	9	6
Furnace CVD	487	239	158	71	36	15	10
Furnace fast ramp	441	217	143	65	33	14	9
Furnace oxide/anneal	285	140	92	42	21	9	6
Implant high current	381	187	124	56	28	12	8
Implant low/medium current	348	171	113	51	26	11	7
Lithography cell	294	145	95	43	22	9	6
Lithography stepper	279	137	91	41	21	9	6
Measure CD	332	163	108	49	25	10	7
Measure film	285	140	92	42	21	9	6
Measure overlay	264	130	86	39	20	8	6
Metal CVD	519	255	168	76	38	16	11
Metal electroplate	268	132	87	39	20	8	6
Metal etch	1153	566	374	169	85	36	24
Metal PVD	591	291	192	87	44	19	12
Plasma etch	1049	515	340	154	78	33	22
Plasma strip	485	238	157	71	36	15	10
RTP CVD	317	156	103	46	23	10	7
RTP oxide/anneal	208	102	67	30	15	7	4
Vapor phase clean	729	358	236	107	54	23	15
Wafer handling	33	16	11	5	2	1	1
Wet bench	474	233	154	70	35	15	10

Solutions exist — Solutions being pursued — No known solution

Source: *The International Technology Roadmap for Semiconductors*, International SEMATECH, Austin, TX, 2003 Edition.

people, and wafer transport/storage represents a very small portion of yield loss (<10%). Particles originate from:

- sputtering/etching of electrode, chamber materials,
- flaking of deposits from chamber walls and wafer holders (chucks, clamps, pins, etc.),
- reaction of chemical species with moisture or oxygen leaks generating solid phase by-products, e.g., gas phase nucleation in tetraethyl orthosilicate $Si(OC_2H_5)_4$ (TEOS) and W-chemical vapor deposition (CVD) processes
- malfunction of filters, purifiers, and other fluid delivery system components,
- abrasion of wafer handlers due to misalignment,
- condensation due to poorly optimized process, pump/vent schemes,
- re-deposition of by-products during wet processing due to non-optimized wet processes,
- stress cracking of deposited films on the wafer surface,
- gas and liquid chemicals.

Process and equipment control are essential for high yields. Typical contamination and defect reduction sources for different processes are briefly described below.

27.2.1 Diffusion and Implant

Besides device design and process integration, trace ion/metal contamination control represents the biggest yield challenge. These impurities are found in silicon starting materials, inert and process gases, liquids, deionized (DI)-water, and ion bombardment debris. Crystalline oxide precipitates, stacking faults, and other silicon structural defects can also impact yield and device performance. During gate oxidation, silicidation and annealing operations, moisture and oxygen entering the reaction chamber (from incoming wafers, system leaks, or reactant gases) can change interface properties, oxide thickness or sheet resistance. Very small particles (<0.1 μm) have the potential of causing early gate oxide breakdown in gate oxides with thickness less than 100 A. The particles are also a major contamination source in polysilicon, nitride, and WSi2 low-pressure chemical vapor deposition (LPCVD) processes due to wall/clamp deposition and subsequent flaking of deposited films. Newer materials for oxides, advanced deposition methodologies, and newer precursor chemicals (typically more complex molecules) for film growth introduce more sources for particle generation or defectivity and require attention at the process introduction to develop reduced-particle or particle-free processes.

27.2.2 Surface Preparation

Most surface preparation processes use acids, bases, solvents, and DI-water. In recent years, chemical quality (with the exception of some solvents) has improved to the point where ions, metals are in the low part per billion to part per trillion levels. However, alkali metals like Ca, Na, Mg, and transition metals like Cu, Fe, Ni can cause gate oxide degradation, junction leakage and device reliability problems even at these low levels. The purpose of surface preparation is to remove surface contaminants or residuals from prior processing. The efficiency of the surface preparation process can determine the resultant level of defects. The improper or inefficient removal of metallic contaminants can result in undesirable enhanced growth or nucleation sites for later film growth which result in defect sites. Such contamination is typically not visible to inline inspection prior to the film deposition. Surface preparation can cause visible defectivity as precipitates or other residue can remain afterwards on the wafers caused by either inefficient rinsing or drying or chemical incompatibility with substrate materials.

27.2.3 Photolithography

Airborne amine contamination is a known cause ofchemically amplified deep ultraviolet (DUV) photoresist profile degradation. The chemical filters and the non-chemically amplified DUV resists are now available to minimize the impact of ambient amines. Airborne hydrocarbon contamination also

deposits on stepper lens causing lens clouding. Particles in photopolymers are particularly difficult to filter and backside wafer particles and film residues introduced from coater tracks and other equipment cause stepper defocus. Improved stepper chuck design to minimize the wafer contact appears to be the best means of minimizing this problem. An overall wafer backside cleanliness program across the entire fab also helps. Wafer backside inspection tools now exist, which can provide insight into the level of backside contamination experienced by the fab.

Besides wafer level contamination, critical dimension (CD) control has a major impact on device speed and yield. Most of the sources of CD loss are from photolithography equipment and related processes. Critical levels such as gate definition, moat patterning and metal 1 push the limits of DUV lithography and etch processes. Depth of focus can be very small resulting in scumming, pattern erosion, and micro-bridging. For sub-0.25 μm design rules, photolithography-related process defects represent a major component of yield loss. Designers use selective size adjusts, phase shifters, optical proximity corrections, and serifs to make it easier to print critical geometries. These additional features on masks can make it harder for mask makers to build defect-free masks. Mask inspection, though typically handled within the lithography groups, is becoming a larger focus of the inline defect inspection groups due to the complicated nature of the newer generation masks. Mask inspection tools result in a dilemma, as a result of reporting the large numbers of defects that ultimately may not result in printable defects. Qualification of the masks is more and more requiring high sensitivity inline inspection of strategically printed wafers to determine the level of printable defectivity and marginality of the photo process with respect to the design elements for the device. Further these masks and mask materials are sensitive to degradation from continued use and exposure to the DUV light sources and must be periodically checked for evidence of this degradation.

In order to push optical lithography to the 65 nm design rule node and potentially beyond, new and innovative resist and resist "assist" products are being introduced which create the need for process defect characterization and optimization. Defect characterization on resist levels has typically been problematic due to the anti-reflective nature of the films. Also inspection techniques with DUV illumination or electron-beam inspection (EBI) that can provide the resolution needed for these design rules may tend to have a deleterious effect of the resist films themselves. Line edge roughness (LER) has also become a critical issue for the 90 nm and below design rules at critical layers as device performance and device metrology can be affected.

27.2.4 Etch Deposition and Oxide or W Chemical–Mechanical Polishing (CMP)

Particles, etch residues and incomplete etches are the biggest yield challenges for interconnect processes. Metal (Al, Ti, TiN) and dielectric deposition, and etch processes generate more killer particles (including a high percentage of large particles greater than 1X design rule) than all other processes combined (with regards to Al-based backend process technologies). Many of the sources of these defects are summarized at the beginning of Section 27.2. Scratches and associated scratch debris from chemical-mechanical polishing (CMP) also cause yield loss (see Figure 27.1). Cu deposition, done by electrochemical (liquid immersion) means rather than physical vapor deposition (PVD), CVD or sputtering has unique defects when compared to the other metals as a result. Filling of high aspect ratios is a significant source of concern but voids are generally not detectable by optical-based inspection techniques. Voids of various kinds (buried and internal to film, large and visible or coalescence of small voids in later processing) are a major concern for Cu deposition and subsequent processing, since smaller voids will tend to coalesce into larger voids with later processing (annealing, stress, etc.). Often the as-deposited Cu film will not exhibit large enough voiding to produce electric signal but with temperature and stress the voids can coalesce to result in electric fails.

Every process and equipment should be monitored and controlled such that contaminants are not formed, or are detected and eliminated before they impact the wafer. However, short product lifecycles and the pressures of introducing new products to market on time have resulted in major design rule shrinks and process changes every 6–12 months. The manufacturing environment is becoming one of

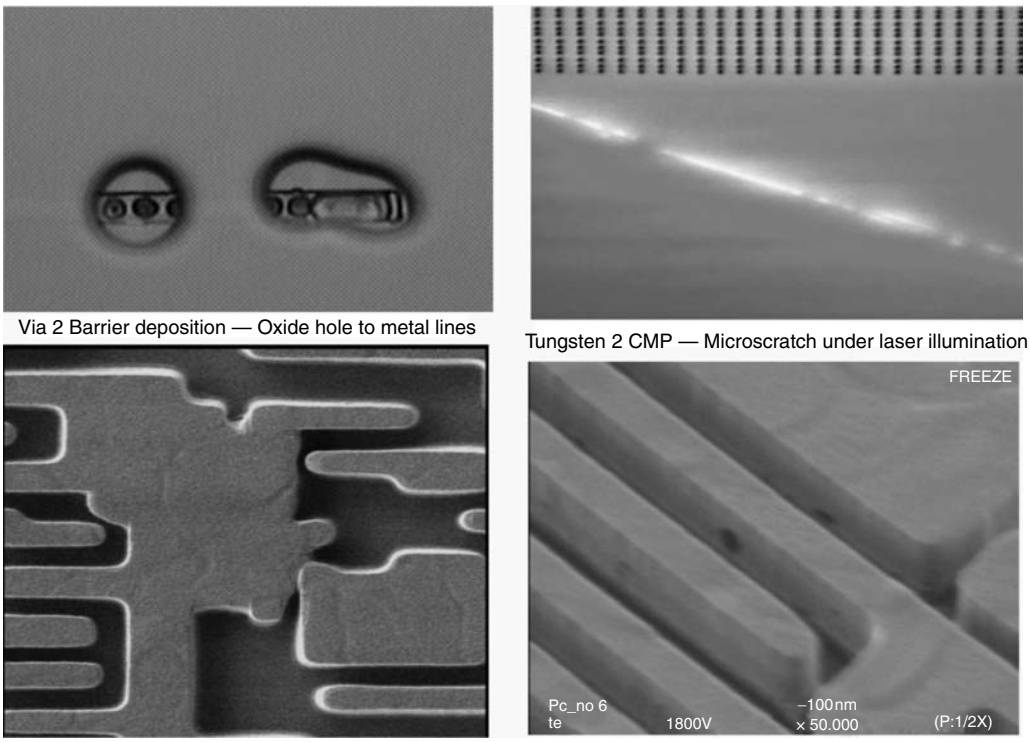


FIGURE 27.1 Random defects seen from etch, deposition and chemical–mechanical polishing (CMP) processes in ≤ 0.5 μm logic fabs.

constant process/equipment change. In this environment, total elimination of random defects is unlikely. Transfer of well-characterized processes from development to manufacturing, and a system to detect and solve yield problems in manufacturing are essential for rapid yield ramps.

27.2.5 Cu CMP and Damascene Multi-Level Metalization

As mentioned previously, the introduction of Cu dual- or single-damascene processing for metal interconnect levels has resulted in new and significant defect mechanisms. These processes have also resulted in significant challenges to the Yield Management groups in terms of strategy and detection of these new defect issues. As interconnect capacitance becomes a more significant factor in the overall speed of device performance the need for lower *K*-value dielectric materials for inter- and intra-level oxides to counter this effect causes integration and defect problems. Some of the new defects introduced by this processing technology include the following (see also Figure 27.2):

- pits or surface voids in the Cu material,
- delamination or "blistering" between films in the dielectric and Cu stack due to stress and adhesion,
- surface residues post Cu CMP cleaning which may becomes sites for initiating delamination or block subsequent via connection to the Cu layer,
- Cu hillocks which are spikes of Cu that rise out of the Cu film after polish and as a result of subsequent heat treatment,
- Cu corrosion where exposure to photon energy results in Cu filaments or Cu extrusions from the trench resulting in shorts,
- incomplete Cu polish resulting in electric shorts, and
- resist poisoning from low-*K* dielectric stack material effluents that interact with DUV resists.

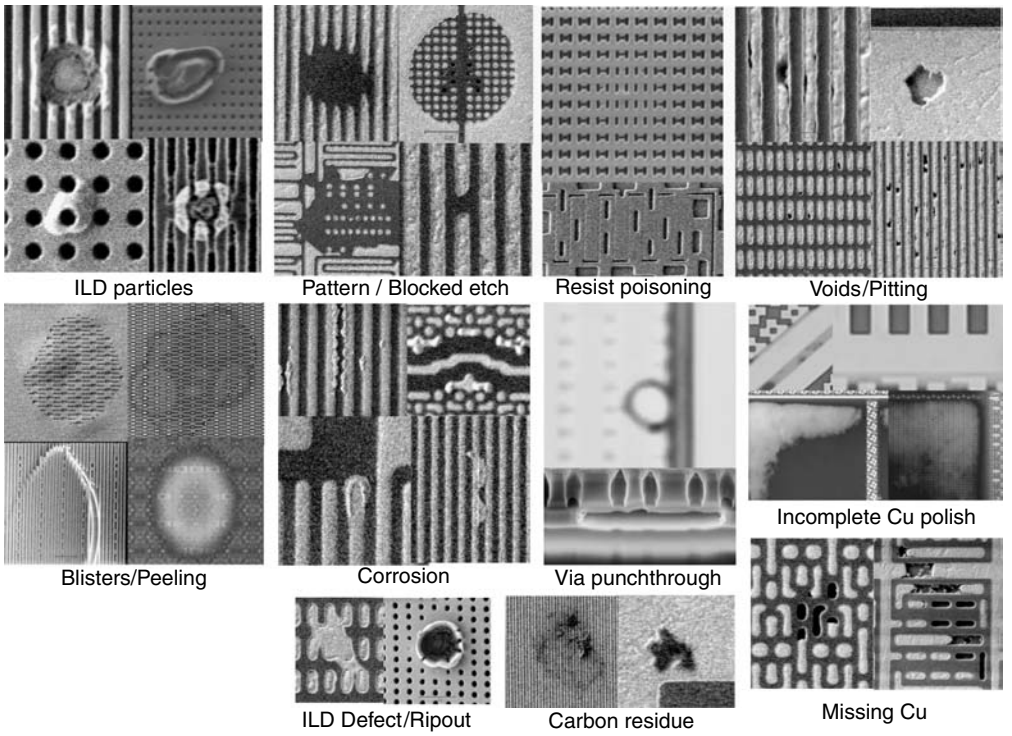


FIGURE 27.2 Examples of typical defects found in the Cu backend with damascene and low-*K* dielectric process technology.

27.2.6 Wafer Edge Engineering

A critical factor in yield improvement has been the proper engineering of the wafer edge region. Film peeling and delamination can be exacerbated at the edge of the wafer (typically the outer 1–3 mm) where the edge bead and exclusion area occurs. This can be due to the improper matching of the films applied to this region (e.g., allowing films to be in contact that normally do not adhere) which occurs due to mismatch of the exclusion zone between photo, etch, and film deposition tools. Some film depositions that do not have edge exclusion can leave residues at the bevel of the wafer that later peel or become loosened and re-deposit back on the active part of the wafer. It has been well-known that the scribe area can also be a source of defects if the scribe process leaves residue. That residue can be moved across the wafer as a result of subsequent processing (typically wet processing). Another edge-related issue is the impact of remaining resist left in the exclusion region at the edge due to improper removal as part of the lithography process. This remaining resist can cause the build up of films or material that may peel from those localized areas. Finally, many sites print reticle patterns outside of the "yieldable" area of the wafer in order to accommodate issues with CMP and other processing. This results in partial die printed. These occur in areas where film and other process uniformity is outside of the expected process window. As a result some of these patterns may be marginal and can break off during later processing and be re-deposited on the inner portions of the wafer causing fails. See Figure 27.3, for examples, of wafer edge regions which can result in defectivity further in due to peeling and delamination. It is important that yield management strategies comprehend the region of the wafer outside of the "yieldable" region which can become a source of significant yield-limiting defectivity.

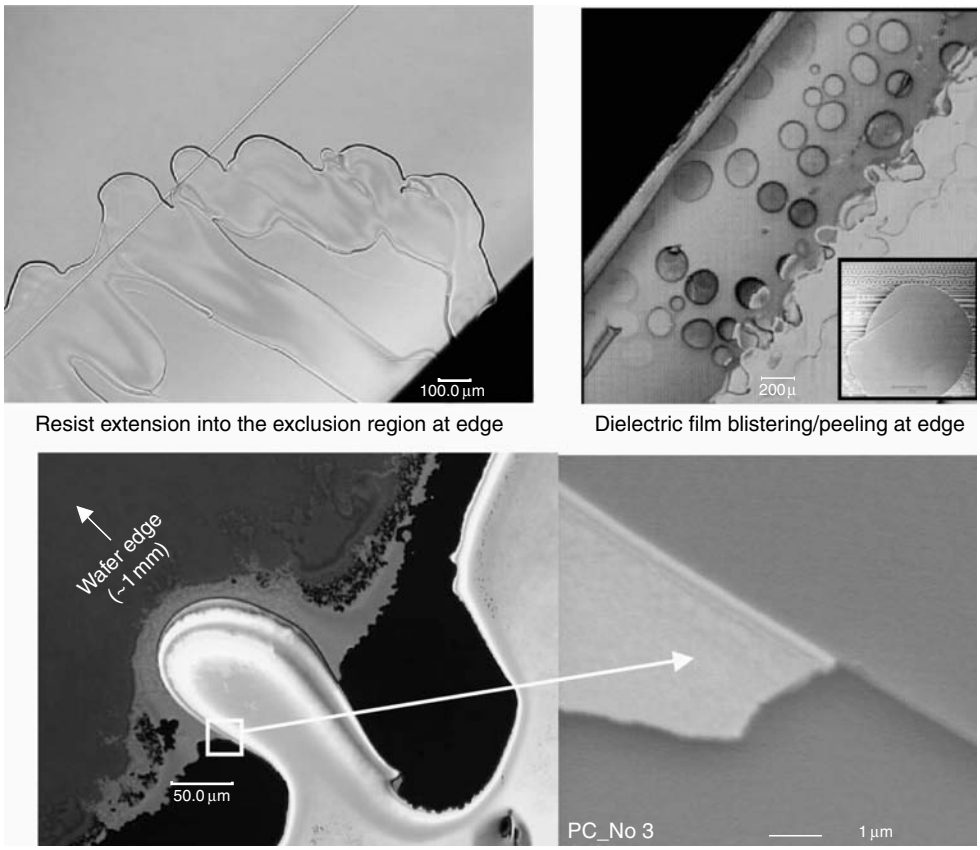


FIGURE 27.3 Wafer edge defectivity examples.

27.3 Yield Management Methodology

The function of a yield management program is to improve yields by predicting, detecting, reducing, and preventing yield losses in the shortest possible time. The intensity of these roles differs depending on the maturity of a manufacturing process. Figure 27.4 shows the stages of a yield ramp. Stage 1 refers to development and early productization, Stage 2 to final productization, Stage 3 to technology transfer and startup in a manufacturing fab, Stage 4 to volume ramp in manufacturing, and Stage 5 to achieving yield entitlement.

During new process development and early productization (Stage 1), the role of a yield management program is to develop and implement an inspection/analysis plan to support process, material and equipment characterization, and diagnose systematic process integration issues. Defect inspection and analysis requirements are defined, and new equipment is evaluated and selected. During this stage, predictive yield models are developed, fab yield entitlement is estimated and unit operation defect reduction targets are set. As previously discussed, a technology node shrink requires baseline defect reduction to achieve similar yields. Hence, a baseline equipment defect reduction program should be initiated in the development/productization fab as well as at the manufacturing site that will receive the new process. Since yields are likely to be low at this stage, metrics of success are improvements in equipment defectivity and reduction of integration-related yield loss.

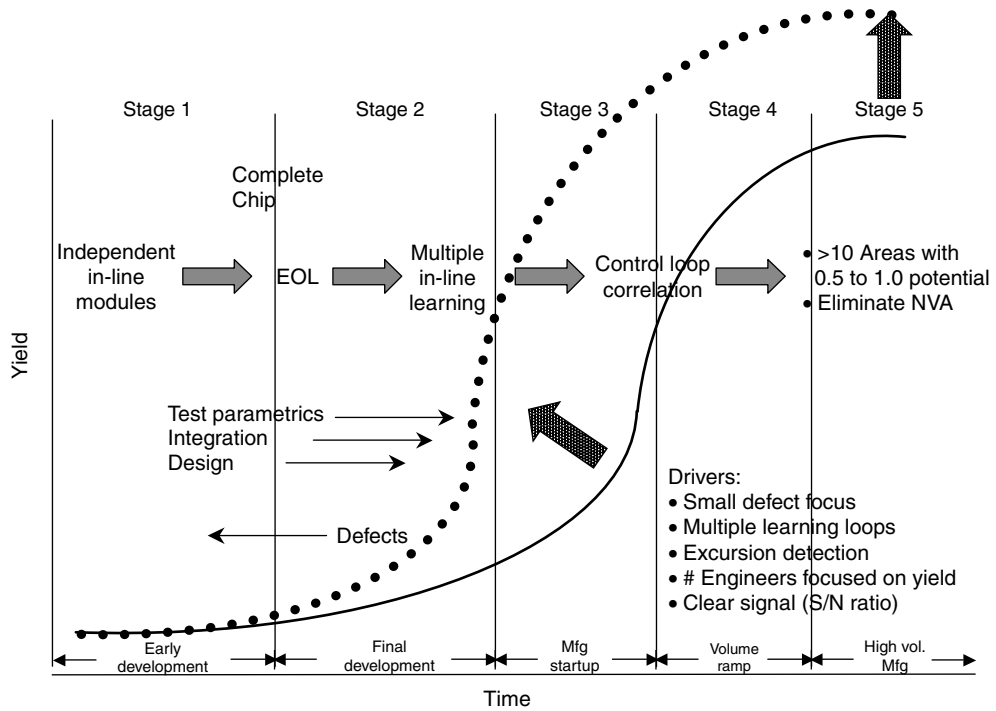


FIGURE 27.4 Stages of yield ramp for new technologies.

As productization progresses to Stage 2, process recipes become firmer, systematic yield losses become well understood (if not solved!) and the focus shifts to demonstration of manufacturability. For yield management, the emphasis shifts from systematic to random defect reduction. A large number of inspection steps at high sensitivity of defect detection are introduced into the process flow to characterize defects from each process segment. A defect pareto is established and baseline improvement is addressed through focused teams. A strategy to deal with excursion control is also implemented. Defects are thoroughly characterized using inspection and analytical instruments. A database of observed defects, root causes, and problem solutions is developed and documented. The yield model developed in Stage 1 is verified and refined with probe data. Stage 2 is also the time when engineers from the receiving fab come to the development fab to be trained on new tools and methods. Metrics for Stage 2 include electrical yield, yield model accuracy, process/equipment defect levels and project schedule.

Stage 3 moves from development to the manufacturing site (or if the sites are the same it transfers ownership from a development team to the manufacturing team). Various companies have developed methodologies (such as "copy exactly") to achieve the same or higher yields in production immediately after technology transfer. Copying equipment and processes facilitates duplication of productization yields in manufacturing. It is also important to copy the same inspection equipment set in manufacturing. This allows for transfer of inspection recipes, defect baselines, and yield models with minimum modifications. Initially the same number of inspection steps and sampling plan should be implemented in manufacturing. Once a baseline pareto is established and critical processes are fingerprinted, some of these inspections can be dropped and the sampling plan reduced. The key metrics for this stage are yield parity with the mother fab (or better), clear in line defect pareto and achievement of baseline defect reduction goals.

In Stage 4, the number of process equipment and the number of lots in line increases dramatically. The objective is to ramp volume and improves on the yields demonstrated in Stage 3, while at the same time qualifying new equipment and processes. If the process is transferred from development to

manufacturing at high yields, the primary objective may be to maintain these yields in the midst of all the equipment additions. A strong, well-trained yield management team with established methods is the key to success. This team must set up and lead crossfunctional yield improvement actions to tackle problems that invariably arise. Major yield gains are often achieved by solving a few yield problems at the top of the defect pareto.

In Stage 5, a fab has completed volume ramp and the yield curve is starting to plateau. Additional yield gains require defect reduction at many points in the process flow. There are no silver bullets or low hanging fruits The entire fab needs to be involved in baseline improvement to see 0.5%–1.0% yield gains. Management commitment to yield improvement is probably the biggest success factor. Excursions have to be detected early so as to minimize the number of impacted lots. Small yield variances and tails in the yield distribution have to be analyzed carefully to squeeze out additional yield improvements. A re-examination of a fab's yield entitlement is warranted at this stage along with cost benefit analysis of the gap between fab yields and entitlement. Additional yield gains may require new equipment or major device redesigns.

At each of the above stages, the following elements form the basis of a good yield management program:

- management priority and motivation
- process and equipment control
- product-based defect detection and analysis
- yield impact prediction
- root cause isolation
- implementation of a solution
- verification of the solution.

There are important differences between process and equipment control and line monitoring methodologies which are part of the product-based defect detection. Principally the process and equipment control methodologies are carried out using unpatterned wafers and the appropriate inspection systems for that purpose. Line monitor methodologies are principally performed on patterned product wafers at pre-planned inspection points in the process flow on the corresponding appropriate inspection tools. It is important that both strategies exist for a successful yield management methodology. Equipment control methods focus on the defects produced by unit processes such as etch, photo, CMP, etc. While unit process defectivity is a critical concern there are significant defect issues that result from unit process interactions usually called integration defects. The line monitoring methods are designed to detect these integration defects, including defects related to topography and patterning issues.

Hybrid methods exist to allow increased capability to use similar inspections to cover both strategies. In particular, some sites utilize product-based tool monitoring. In this methodology, lots are sampled for inspection both before and after a unit process. The added defectivity is considered the unit process defectivity. This method, however, suffers from several issues of its own:

1. replacing lower cost inspection (unpatterned wafers) for higher cost inspection (for this purpose the typical inspection tools that would be used would be of medium cost comparatively),
2. issues with sensitivity for some processes before and after processing (e.g., film deposition may highlight previous defects that were not previously detected, thus causing an errant high adder defects),
3. logistical issues with sampling lots appropriately to cover the entire process tool base for a particular unit process, and
4 increasing the cycle time of production lots by implementation of more rigorous and more frequent inline inspections.

One of the advantages of the hybrid too monitoring method is particularly attractive to 300 mm substrate wafers fabs in that the use of pilot wafers and associated costs can be dramatically reduced.

Another important advantage of the hybrid method is the ability to correlate tool defectivity with defectivity on product. Whether the increased costs associated with the increased logistics, cycle time, and inspections offsets the savings in wafer costs must be evaluated for each fab. Furthermore the ability to successfully identify defects of interest and pinpoint the source should be a consideration.

27.3.1 Management Priority and Motivation

Every fab has multiple priorities, such as cost, cycle time, yield, and wafer output. The rate of yield improvement is closely tied to the priority assigned by management to yield. Progressive fabs set clear goals and have visible indices to track yield and offer incentives for achieving or exceeding goals. Employees at all levels are encouraged and motivated to improve process control and reduce defects. It cannot be emphasized enough the importance of management support at all levels to prioritize yield improvement efforts. Yield improvement is a multi-disciplinary effort requiring the cooperation of all groups to achieve lasting success. Most managers would agree that yield improvement is an important and vital need but many do not translate that into actual goals since such efforts are not perceived to be under their direct control. Such goals appear to be ill-defined with regards to their typical daily activities. A culture where yield improvement activities is emphasized and rewarded must be developed and continually encouraged since the work is often in conflict with other goals such as improved lot cycle time.

27.3.2 Process and Equipment Control

The mechanisms of particle formation in process equipment are different in each of the cases discussed in Section 27.2 and a fundamental understanding of the physics and chemistry of many of these mechanisms is lacking. Some of these sources, such as condensation or gas phase nucleation can be systematically eliminated through process recipe optimization and elimination of system leaks. Others, such as flaking can be minimized through regular chamber cleaning, use of parts with high surface area or use of special procedures like "PVD pasting". During Ti, TiN sputter deposition, chamber walls are "pasted" with Ti after each wafer is processed to "glue" any flakes to the chamber wall. Regular preventive maintenance can also go a long way to minimize particle events. In many instances the only recourse is to clean wafers after a particularly dirty process. For example, brush or megasonic scrubbers are commonly used to clean product wafers after sputter deposition or chemical mechanical planarization.

Regardless of the origin, it is important to know the quality of process/equipment on a regular basis. Factories monitor particle levels of all equipment using unpatterned pilot wafers once per shift or once per day. Wafers are scanned through an unpatterned wafer inspection tool before and after a pilot run to get the number density of particles added by the process tool. To obtain good readings, it is important to process pilot wafers under conditions similar to those used for product wafers. Far too often, factories monitor equipment without running the process, i.e., they do not pass process gases through the chamber or do not turn the radio frequency (RF) power on. Under such conditions, these measurements are poor, or completely irrelevant indicators of the health of process or equipment. Pilot wafer-based equipment monitoring and control is the most common approach to determining if a piece of equipment is in a "ready" state for production. They are used to qualify a new process, bring a tool to production readiness after a maintenance operation, or for routine checkup. Particle measurements are plotted on control charts and standard statistical process control (SPC) rules are applied to determine the next step.

Since equipment can be shut down based on pilot wafer readings, it is important to establish credibility of these measurements by periodically correlating them to product wafer measurements using patterned wafer inspection equipment. With higher throughput patterned wafer inspection equipment, there is a trend to do away with unpatterned wafer inspections altogether and replace them with product inspections. This reduces pilot wafer costs and equipment down time associated with monitoring. It also eliminates the need to correlate pilot and product measurements. It is also needed when monitoring

processes that are prone to an increasing frequency of wafer-to-wafer defect excursions as a tool-related failure mechanism. In this case, monitoring defectivity of one or two wafers on a daily basis does not give a true picture of the health of the tool. Figure 27.5 shows defect counts of 12 unpatterned tool monitor wafers run sequentially in a process tool in poor condition as seen in the defect excursion rate. Single wafer dielectric deposition tools with in situ cleans linked to wafer count or deposition thickness are a prime example of such a tool. In this case, unpatterned tool monitoring capability is only needed for diagnostic and post-maintenance qualification before a return to service. As a result, it is unlikely that unpatterned wafer-based equipment monitoring will disappear in the near future and also patterned wafer inspection machines are still 4–6 times more expensive and 3–10 times slower.

Particle wafer monitors are limited in their usefulness because they only provide a snapshot of process quality a few times a day. Ideally, we would like sensors that monitor the process in real time. Such in situ process sensors (ISPMs) have been demonstrated to work in only a limited number of process and equipment combinations.

27.3.2.1 In Situ Contamination Monitoring

Real time contamination monitoring of process equipment is desirable because it reduces the time to detect an excursion or baseline shift. Such sensors have been used for many years to monitor particles, moisture, and oxygen in bulk gases, particles and total organic carbon (TOC) in DI-water, and particles in liquid chemical distribution systems. Sensors for use on process equipment have had limited success. In situ particle sensors have been successfully applied to monitor some etch, CVD, diffusion, ion implant and wet cleaning equipment. These sensors are installed in the exhaust line, downstream from the process chamber. Residual gas analyzers have been used for process diagnostics on a few processes, but their application has been limited by their size and the expertise required to understand complex spectral information. When they work, ISPMs provide rapid information about contaminants generated during actual process conditions.

Most in situ particle monitors detect particles using scattered light. Figure 27.6 is a schematic of one such sensor. A laser and set of optics create an intense-focused beam of light projected through an area particles travel. Most in situ sensors use laser diodes as the optical source because they are small, powerful, reliable, and relatively inexpensive [8]. As particles traverse the beam, light scatters in short pulses to one or more detectors. The detectors, typically photodiodes, convert the light pulse to an electrical signal that is counted over time. The small, modular design allows flexibility in locating the sensor in the exhaust line near the process chamber of the tool. Since ISPMs are near the process chamber

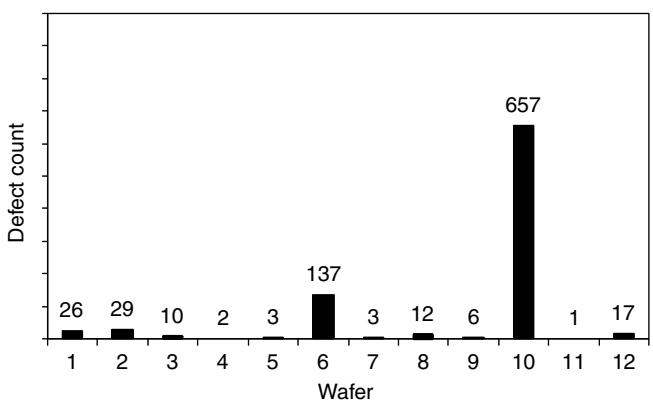


FIGURE 27.5 Defect counts on 12 unpatterned tool monitors run sequentially in a process prone to excursions.

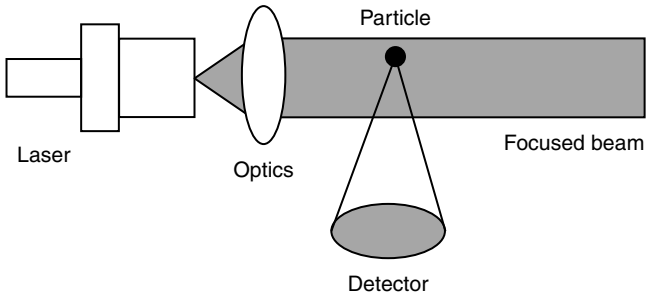


FIGURE 27.6 Schematic of an in situ particle sensor.

where they are exposed to vibration, corrosive chemicals, electrical noise and optical interference, the sensors must be robust. The application requires a high signal-to-noise ratio and the sensor often uses optical filters to reduce interference.

An example of a successful particle ISPM application is on AMAT P5000 tungsten CVD systems (see Figure 27.7). Menon and Grobelny [9] reported good correlation of ISPM data to in line wafer defect data and metal short fails, document particle "signatures" for excursions, and verify that the NF_3 clean step is the main source of particles in the process. The ISPM data in Figure 27.7a indicated a shift in the particle performance of the W-CVD chamber nearly 3 weeks before the pilot wafer particle check failed (Figure 27.7b). Note in Figure 27.7b that the particle levels on the pilot wafer check are not statistically different before and after the chamber clean. The pilot wafer particle check did not indicate a particle problem with the W-CVD chamber until there was a catastrophic flaking of particles (the chamber was then cleaned). In Figure 27.7a, the ISPM data show a distinctive increase in particles well before the chamber was cleaned.

A major deficiency of current ISPM sensors is that they do not "plug and play". Each equipment application requires customization. Because they are installed downstream of the process chamber, they detect particles only if pressure and fluid transport streamlines are conducive to moving particles from the chamber to the detection point. In low pressure processes like sputter deposition, gravitational forces are much larger than drag forces, causing particles to deposit in the chamber rather than being transported through the exhaust line. Takahashi and Daugherty [10] have published an extensive review of this technology, its advantages and limitations. In situ process sensors, however, address only a portion of the defectivity inline that affects yield: that related to fall-on particle contamination. Defects due to process integration, for instance, would not be captured by this technique. In the past few years ISPMs have been integrated into the semiconductor process equipment tools where they are applicable and are now part of the standard control system.

Finally, one of the drawbacks of ISPM monitoring is the relative lack of ability to correlate to yield-limiting defectivity initially. There are multiple stages in some processes that produce particle type effluents but are not causing particulate contamination on the wafer surface, at least in such a way that causes yield loss. The ability to directly demonstrate correlation of a detected defect problem to yield loss is one of the significant tools in yield management. By drawing such causation the yield engineer is able to more effectively drive management and engineering resources to eliminate the defect problem. As defect monitoring methods move away from patterned wafer analysis such causation becomes more difficult to ascertain. In the case of ISPMs there is a need to focus on those stages in the process which ultimately produce the yield-impacting defectivity. Typically such correlations are determined by studying the correlation between the ISPM readings from various process stages and the impact on wafers in the system at the time. Therefore, new processes will need such characterization in order to successfully implement ISPM monitoring.

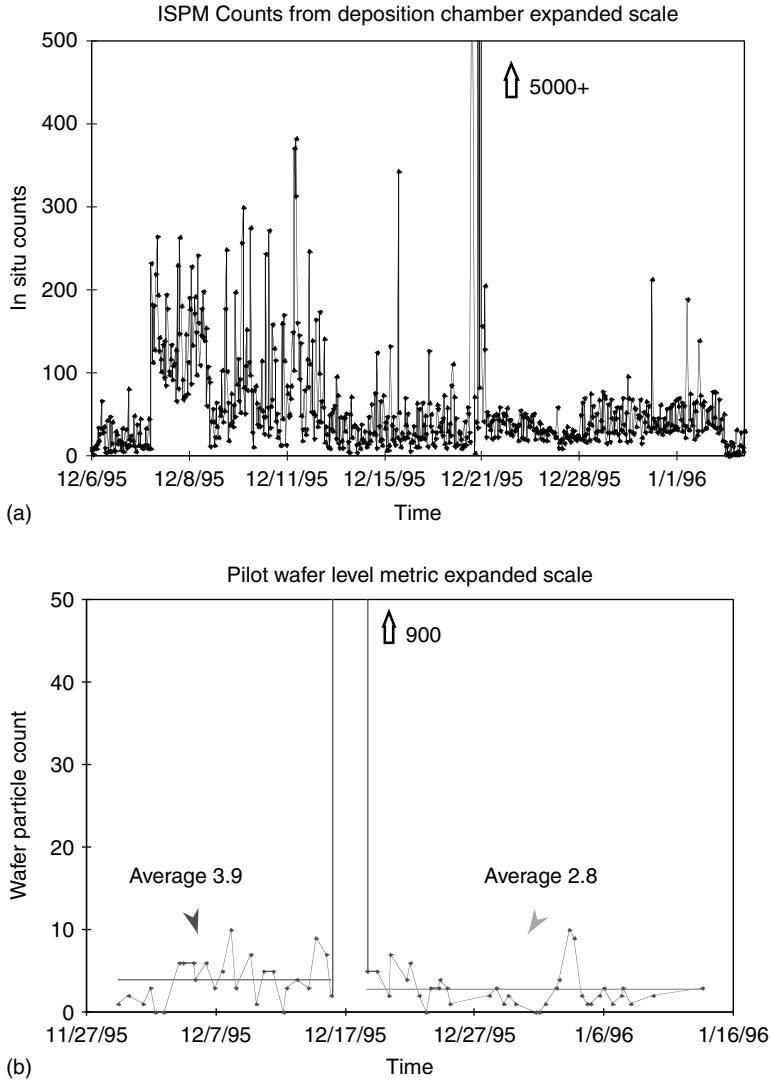


FIGURE 27.7 Results from an in situ process sensors (ISPM) application for tungsten chemical vapor (CVD) deposition in an applied materials P-5000 chamber. (From Menon, V. and Grobelny, M., *AVS Symposium*, San Jose, CA, 1997.)

27.3.2.2 Unpatterned Wafer Monitoring

Since ISPMs address a limited number of applications on process tools there is still a heavy reliance on particle measurements on test wafers to qualify, if not monitor process equipment for defectivity. The majority of these measurements is done using unpatterned pilot or test wafers (unless the site has opted for a tool monitor method on product wafers). A variety of inspection systems that are design to inspect wafers with blanket films or no films and have no pattern are available. These typically will use laser-based scattering inspection methodology. Due to the lack of pattern "interference" with the inspection, these systems can achieve relatively high sensitivity to particles, scratches, pits, etch blocks, other defects, and even to surface roughness of the films themselves. The inspections are typically fast in that a 200 mm wafer can be inspected in about 1 min or less. Most patterned wafer inspection tools can also perform unpatterned wafer inspection as well but typically at a reduced throughput.

Since unpatterned wafer inspection tools cost much less than the patterned wafer tools, such application is not considered appropriate unless the type of illumination and detection available on the patterned wafer inspection provides an added benefit.

Unpatterned wafer inspection tools are typically sensitive to the film or film stack on the wafer and thus need to be calibrated or optimized to obtain maximum sensitivity. A unique aspect of these systems is that they can be calibrated to traceable known size National Institute of Standards and Technology (NIST). These standards are typically polystyrene latex (PSL) spheres that can be obtained in several calibrated sizes. The ability to detect these spheres of differing size bins on particular film stacks allows a measure of sensitivity that is traceable to NIST.

However, the defects that are encountered in semiconductor manufacturing are mostly not spherical like the PSL spheres and are of a wide variety of materials that differ from PSL. Due to the fact that the illumination light interacts differently to the defects of interest, the sensitivity to these defects may be significantly different than that achieved using PSL spheres. In addition, the NIST standard size PSL spheres are currently larger than the size range of defects currently of interest and the calibration to these sizes may not be particularly relevant to existing semiconductor manufacturing needs. Achieving sensitivity to defects of interest (90 nm and below design rules) is more often accomplished by optimizing the signal-to-noise using the variety of on board analysis capabilities than the PSL sphere technique since the latter requires special equipment and techniques to properly produce calibration wafers.

It is important that scanning electron microscopy (SEM) review of the defects detected on the unpatterned tool monitor be performed at initial monitor setup and periodically to aid in trouble shooting tool defect excursions. The SEM review at monitor setup is used to build an understanding of the defect types that are present on the monitor. This is important for several reasons. One is to make sure that the monitoring is detecting the defects of interest known as the defect of interest (DOI). The DOI are the defects which have impact on yield and product reliability as determined through the defectivity seen inline and from end of line failure analysis (FA). This method has also increases the fundamental understanding of the process-related defectivity allowing a clearer differentiation from integration-related defect mechanisms. Setting up tool defect monitors solely on indicated defect size relative to PSLs often results in monitoring of nuisance rather than the defects of interest or missing a defect type entirely which is perceived as the noise floor of the wafer.

27.3.2.3 Wafer Backside Inspection

As lithographic depth of field process windows decrease with shrinking design rules the impact of wafer backside contamination on frontside defectivity increases [13–15]. Particles, residues, or pits and damage on the backside of the wafer can be correlated to lithographic "hot" spots or out-of-focus spots. Backside defectivity can also result in localized defectivity in etch or other processes if there is a lack of contact between the substrate and the tool chuck which causes improper heat distribution or if severe enough can cause arcing in severe enough cases. Finally, backside particles and residues can be passed on to the wafer frontside during subsequent processing. Tools specifically designed to inspect the wafer backside which have been available for several years to address this need however techniques to do so using prior unpatterned wafer tool sets have been in place, usually requiring the sacrifice of the frontside of the wafer. Some manufacturing facilities include wafer backside measurements to qualify and monitor tools or processes which are historically likely to cause wafer backside contamination. An example of wafer backside inspection capturing a defect which results in frontside defectivity is shown in Figure 27.8.

Higher backside sensitivity and less noise can be expected with 300 mm wafers when compared to 200 mm or smaller wafer sizes particularly at the early stages of the process since 300 mm wafers will be polished on the backside. As the wafers proceed through the process, though, the results of the processing and handling add significant backside defectivity which is not necessarily relevant to front side defectivity and ultimately yield. Analysis of backside wafer maps is largely based upon "signatures," that is, spatially-distributed patterns of defects in recognizable shapes due to the high defect counts typically encountered. Applications and data analysis of wafer backside inspection defectivity and patterns are still in a relatively

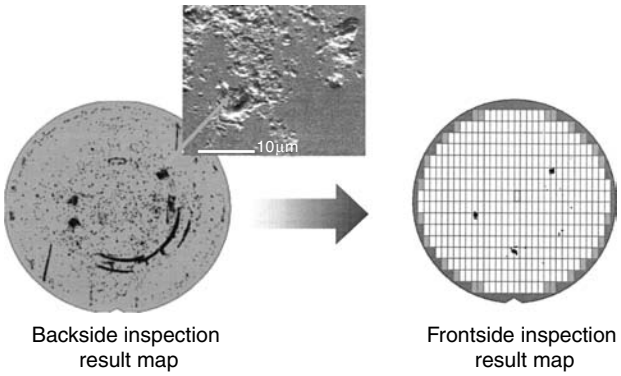


FIGURE 27.8 Example of a hotspot induced by backside defects.

early stage of development. These analyses are aimed at reducing the high defectivity counts and patterns to identify the backside defects and defect signatures of interest.

27.3.3 Product-Based Defect Detection and Analysis (Line Monitor)

We have always relied on electrical test, with associated electrical and physical analyses, as a definitive measure of yield loss. However, cycletime to detect a defect problem can be 30–60 days (see Figure 27.9). For advanced flows up to 11 layers of metal the cycletime can be in excess of 60 days. In the last 15 years, in-line defect detection equipment has matured to a point where they are deployed at critical points in the process flow to characterize process defectivity (I.L.M. in Figure 27.9). This has reduced the time to detect random defects to a few days or even hours.

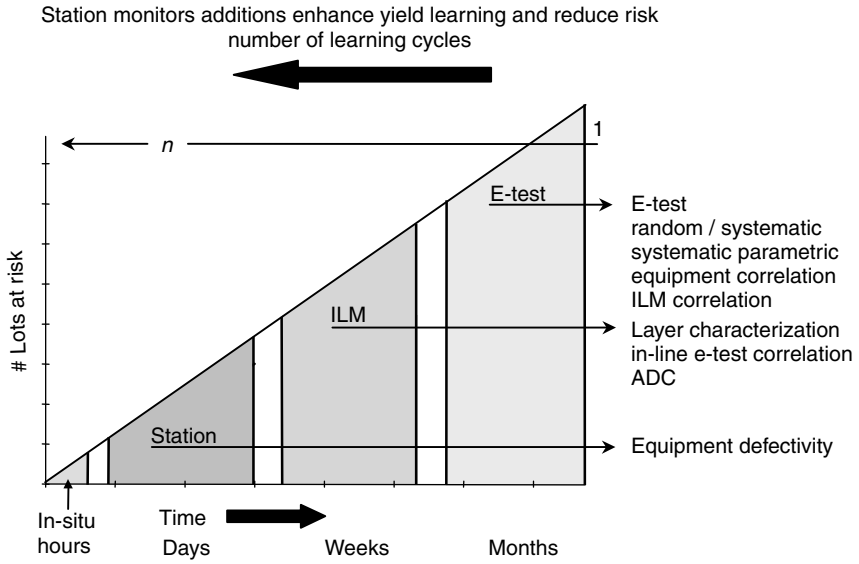


FIGURE 27.9 Relative timescales to detect defect problems in a factory. (From Menon, V. and Grobelny, M., *AVS Symposium*, San Jose, CA, 1997.)

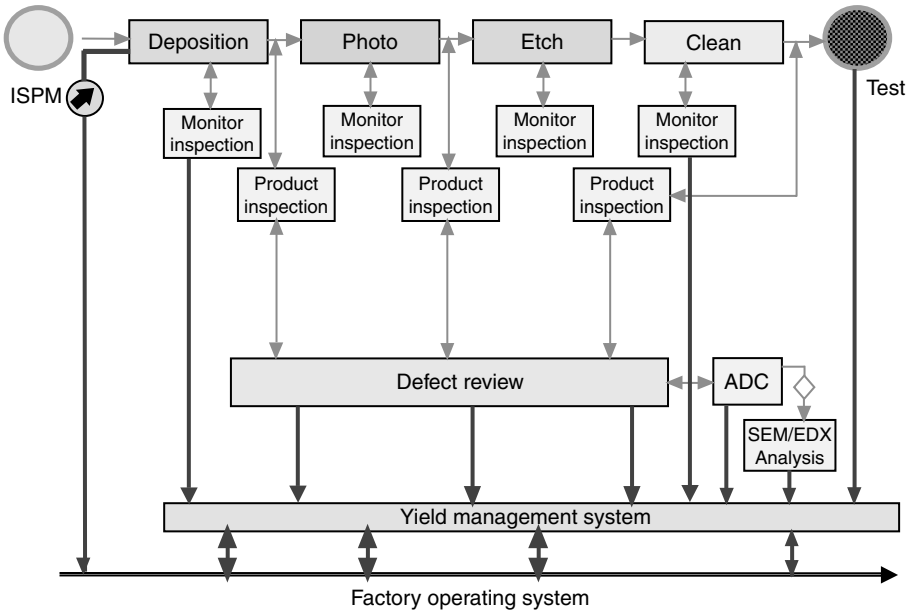


FIGURE 27.10 Typical tool set and inspection locations for yield management. (Thicker lines indicate data flow; thinner lines indicate wafer movement.)

A typical tool set deployment for defect detection and analysis is shown in Figure 27.10 and consists of inspection, review, and analysis tools, and a yield management system. Patterned wafer inspection equipment are used after deposition, etch or photolithography operations, while unpatterned wafer inspection equipment are used for equipment monitoring. These inspection tools provide size, number density, and coordinate location of detected defects. In addition, some of the current inspection tools have the capability to allow for classification of the defects with some added time from training the inspection tool based on defect features Automatic Defect Classification (ADC). The inspection tools may also provide optical images of selected defects. Optical and SEM review equipment are used to view and classify defects which can augment the inspection tool capability or be used independently if the inspection tool classification capability is too limited for the defects in question. Defect review and classification provide visual confirmation and additional information (shape, texture, color, location above or below a process film etc.). A yield management software/hardware system collects and analyzes data from these inspection and review tools, and correlates them to electrical test, parametric test, and factory computer integrated manufacturing (CIM) data (such as equipment ID, process recipe, etc.). One should also not underestimate the need for defect images to provide.

The objectives of an optimal inspection plan are to provide a baseline defect density profile of the process and to detect defect excursions immediately after they occur. A sound sampling strategy will minimize the detection delay after a defect excursion thus minimizing material at risk. Obviously, one needs to accomplish this at a reasonable cost of inspection. Figure 27.11 depicts questions that need to be answered before an in-line defect detection plan can be put in place [2]. Where should one inspect? What percentage of lots should one inspect? How many wafers per lot? What percentage of the wafer area? At what defect size sensitivity?

Typically, 2–3 product wafers from 50 to 100% of all lots are inspected on patterned wafer inspection tools. The same wafers are inspected at different points in the line, allowing for calculation of cumulative number of defects in the line as well as partitioning of defects from different process segments. A 0.25 μm logic or memory process flow may have 25–40 inspection points depending on the number of deposition,

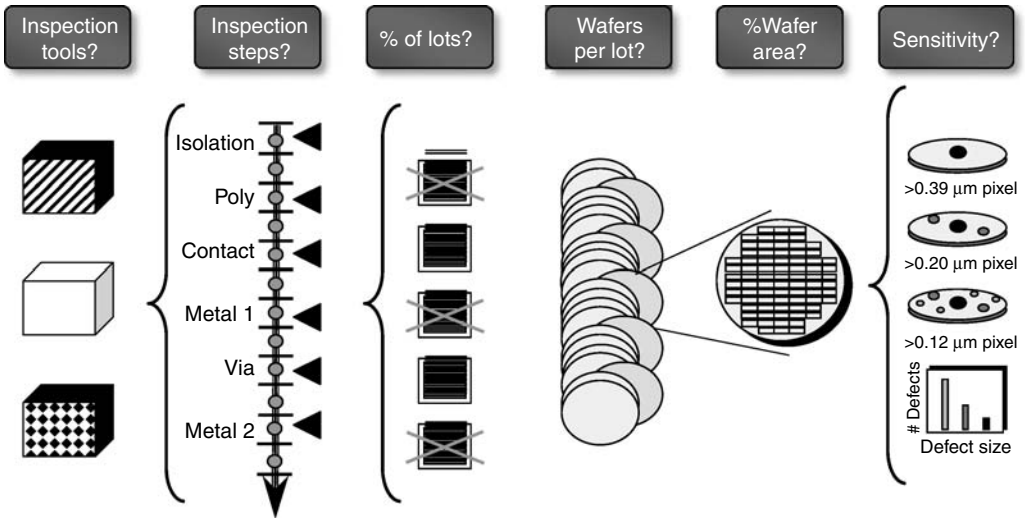


FIGURE 27.11 Factors to be considered while developing an in-line defect sampling plan. (From Akella, R., Jang, W., Kuo, W.W., Nurani, R.K., and Wang, E.H., *KLA Yield Management Seminar*, Santa Clara, CA, 1996.)

CMP, etch, and photolithography steps. For 65 nm logic process technology flows, the inspection log points scales roughly with the increase in number of metal layers from typically four to seven or more. The inspection sensitivity is process dependent, and the user has to balance the need to detect as many true defects as possible with the need to keep wafer throughput high and false alarms low. Figure 27.12 is an example of an in-line defect pareto. It shows the number of defects detected at each inspection point after any previously detected defects that have been subtracted. Essentially, it provides for "loop" level defect density. The risk of not detecting a defect excursion increases as the sampling rate drops. In studies by Intel [3] and Advanced micro devices (AMD) [4], lot to lot variability was found to be much higher than wafer to wafer variability. Hence it is better to inspect more lots than more wafers per lot.

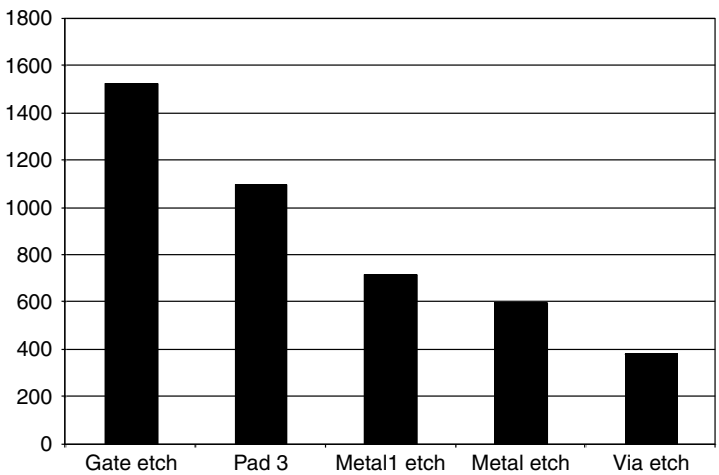


FIGURE 27.12 Example of an inline defect pareto.

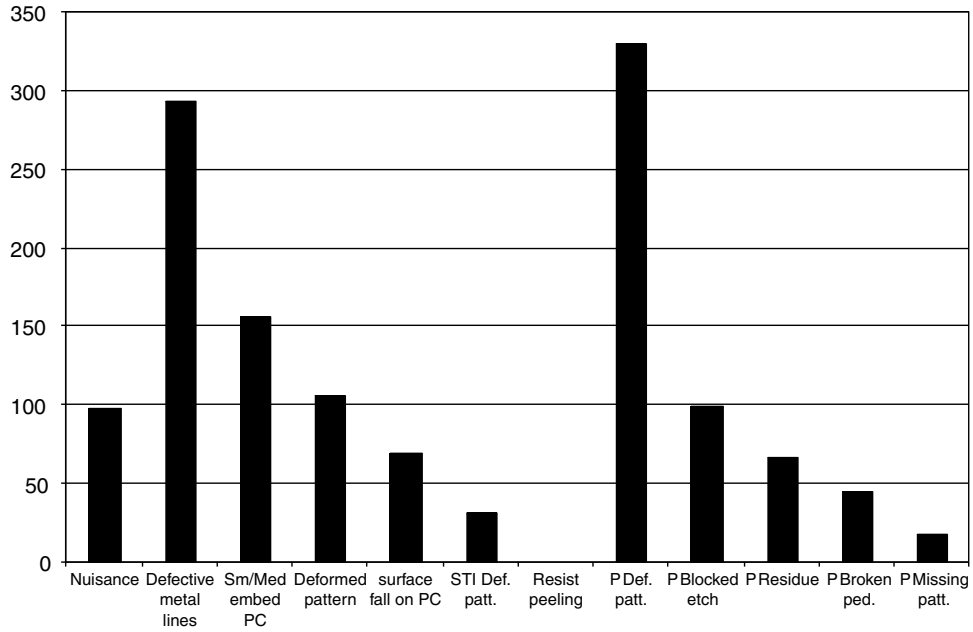


FIGURE 27.13 Manually classified pareto for a single inspection point (metal 1 etch).

The goal of defect review and classification is to provide an additional information on specific defects that allow for rapid root cause identification. This represents the best way to segment loop level defect density to individual process steps within the loop. Each patterned wafer inspection will have a classified defect pareto. Different companies use different approaches to defect review and classification using optical microscopes and SEMs. Some review all defects detected in-line, while others review only "out of control" wafers or a certain percentage of defects on each wafer. Manual review and classification is done by the trained operators who refer to a database of defect pictures and a classification "naming" code. For example, the number of large defects will be separated from small defects, particles will be separated from etch residue, or defects lying on top of the pattern will be separated from those under the pattern. Figure 27.13 is a typical pareto of classified defects.

As mentioned previously, ADC algorithms have been incorporated into patterned wafer inspection tools so that the classification can be done automatically after each inspection. This has resulted in a big improvement in the accuracy and precision of classification over that of human operators. Figure 27.14 shows the performance of an ADC system that is built into a patterned wafer inspection tool. Accuracy of classification is determined relative to an "expert" human. Automatic defect classification accuracy is reported to be over 75%. In contrast, accuracy of human operators across multiple shifts is typically below 50%. Another advantage of ADC is the reduction in time to results. When classification is done manually on a separate optical review station, additional queue time is introduced between inspection and classification. In the study by Bennett et al. [5] ADC was over three times faster than manual classification (Figure 27.15). This technology is fast gaining acceptance as a valuable tool for yield management.

Original ADC algorithms required the collection of a sample of defect images after the completion of the initial inspection. This is due to the need for higher resolution imaging than typically used for the inspection itself. Unfortunately this leads to a trade-off with throughput and inspection tool capacity as it adds time to the overall inspection. Newer algorithms have allowed for use of the information available from the defect during the inspection which allows for classification of all defects without significant

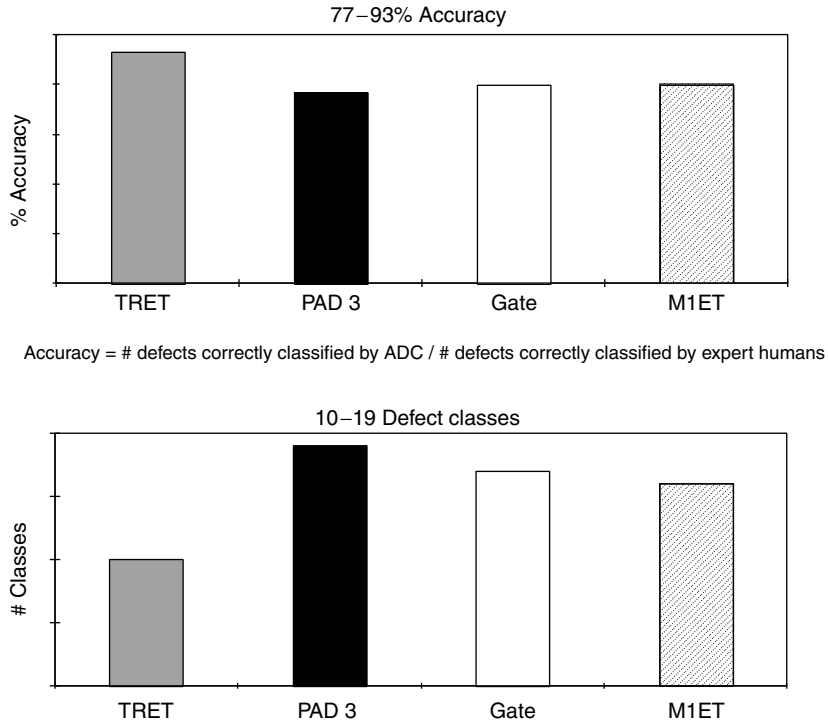


FIGURE 27.14 Typical accuracy from an automated defect classification system applied to a 0.25 μm process flow. (From Bennett, M., Garvin, J., Hightower, J., Moalem, Y., and Reddy, M., *KLA Yield Management Seminar*, San Francisco, CA, 1997.)

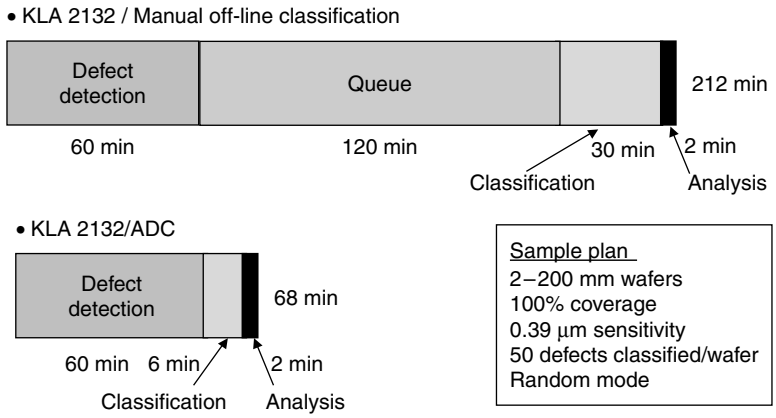


FIGURE 27.15 Comparison of (the time required) classification manual vs. automatic defect classification (ADC). (From Bennett, M., Garvin, J., Hightower, J., Moalem, Y., and Reddy, M., *KLA Yield Management Seminar*, San Francisco, CA, 1997.)

added overhead in the inspection process. Since this is typically performed at a lower resolution imaging the ability to categorize into detailed bins is limited. However, since categorical information on all defects can be obtained during the scan this allows for different applications such as nuisance filtering and directed post-inspection review sampling. With nuisance filtering, a nuisance-defect type that may be hard to remove by other filtering methods which can be classified and ignored in the final defect reporting. Directed post-inspection review sampling means that the post-inspection review sampling (as typically used for optical or SEM review) can be weighted towards defects of particular "coarse" defect type bins identified from the initial inspection to provide finer detail and to highlight defects that have a stronger potential to be yield problems. Such sampling has be done based on defect size distributions in the past but now similar algorithms can be applied to a wider range of attributes. Typical sampling has been done in the past using algorithms that select defects based on a random sampling.

The primary challenges of defect inspection for current process technology are, in general, similar in nature to prior recent generations: (1) ability to detect smaller defects; (2) faster inspections with adequate sensitivity; (3) ability to detect defects of interest (not just smaller but low contrast or low profile defects); (4) ability to screen out noise sources; and (5) ability to quickly and effectively detect defects in high aspect ratio inspection (HARI) structures [23,24]. Some of the challenges in defect inspection and analysis are in the areas of small defect detection (<0.1 μm), fast detection of residues and partial etches at the bottoms of contacts or isolation trenches, and the ability to separate signal from noise (due to surface roughness, metal grains, high reflectivity films, 3D topography, etc.).

A variety of inspection systems and techniques are available for the yield engineer to apply to a given problem to detect DOI. A yield management strategy includes the application of the appropriate inspection technology to a given process level to insure adequate ability to capture all relevant defects while minimizing the detection of nuisance events. The following sections review the automated inspection technologies available to date for application to the semiconductor process. The following review should be considered as generally applicable and any detailed information on operation and features should be requested from the system manufacturer.

27.3.3.1 Brightfield Inspection

In the early days of semiconductor manufacturing, manual optical microscope inspections were a primary means for detection of the manufacturing defects inline. Microscopes would typically have two modes of operation: (1) brightfield; which is the use of the primary reflected light from the illuminator to view the wafer surface and (2) darkfield which placed a filter to block the primary reflected light and instead viewed only the backscattered (off-normal) light. The two primary inspection systems in use today use the same basic principles for automated inspection. Due to the need for high resolution on 90, 65 nm and smaller design rule features the brightfield systems have had to migrate to UV and DUV illumination. However, caution may need to be exercised with regards to the types of films being exposed to DUV illumination in particle since film damage may occur.

In addition to the illumination source wavelength reduction, these is a need for higher magnification for detection of smaller defects. Typically the magnification is referred in terms of "pixel size" since the image is digitized and the pixel size corresponds to one side of square region of the wafer that maps into a single pixel of the detector.

The brightfield systems in use today generally will collect digitized images of similar sites in adjacent die (or analogously similar sites in adjacent cells in a repeating memory or array mode). A particular pixel is assigned an appropriate grayscale value depending on the amount of light that is reflected from that region. Simply speaking these images are compared by image subtraction producing a difference image. In the ideal difference image the background is dark, while an area affected by a defect in one of the images (usually three images or locations are compared) appears bright due to the pixel grayscale difference.

In addition to the different illumination sources (white light, UV, and DUV) and pixel sizes (down to 0.12 μm) the current brightfield platforms offer a variety of other options for inspection including a darkfield-like transformation of the digital image and a variety of digital "filters" and screening methods.

Best known methods for brightfield inspection application include post-etch and post-photolithography processes which should be considered as an initial guideline for best application of an inspection technology.

27.3.3.2 Darkfield Inspection

Darkfield inspection technology typically refers to systems that are based upon laser light for illumination of the substrate surface. This illumination can be applied at an angle to the substrate surface or with normal incidence to the wafer surface. Due to the nature of laser illumination the light will be monochromatic and of high intensity. As a result the detection of light for defectivity occurs at angles off the 0th order reflected path. Thus these systems will collect scattered light which can come from both the pattern on the wafer and any defects. The high intensity is needed for sensitivity as the scattered reflected light from a defect or particle may be substantially less than the 0th order reflected beam light, however, too high of a laser intensity or power could potentially damage the film.

Detection and sensitivity are largely related to the ability to screen out the scattering due to the pattern and noise and highlight the scattering signal due to defects. Fortunately pattern features will tend to scatter in certain directions (depending on the angle of illumination, wavelength, and other factors) which allows for good sensitivity to defect scattering at other directions. Also special spatial filtering can be applied to array type structures since the arrays structures scatter in well-defined patterns based on the array periodicity. This special filtering can allow even higher sensitivity in the arrays. The basic principle behind this is the fact that light scattered from a periodic structure will be diffracted into well-defined spatial nodes due to constructive and destructive interference effects. The size, number, and shape of the nodes will depend on the periodicity of the array and the angle of the incident light. This diffraction pattern can be considered the Fourier transformation of the array. By creating a spatial filter to block these nodes, one can filter out all of the light scattered due to pattern in these structures. The remaining scattered light that is detected is due to defects with a much high signal-to-noise capability than can be obtain from non-array type pattern.

Depending on the illumination angle which varies from normal incidence to high angles on these types of tools the sensitivity to types of defect can vary. Systems which employ off-normal angles of incident light will tend to have higher sensitivity to defects that either extend out of the current process layer of the wafer (e.g., particles) or extend below the surface of the current process level (e.g., pits or holes). Systems that use normal incidence illumination will tend to have higher sensitivity to planar defects such as pattern issues or blocked etch. More advanced systems attempt to provide a combination of both the normal and the angled illumination to attempt to provide a broader range of defect type sensitivity. In the past, darkfield tools have primarily been utilized to inspect post-film deposition or post-CMP processes since the substrate at those layers is relatively flat and uniform and the primary defect of interest are particles or pits and scratches. With the advances to darkfield technology in the past few years the applications of this technology have expanded beyond these levels to include some of the more traditional levels associated with brightfield tools. For a particular application it is recommended that a verification and comparison be made to determine the trade-offs in defect capture with each inspection technology.

27.3.3.3 Electron Beam Inspection

To meet the need of the continued reduction in feature size and the need to quickly determine the extent of "non-visual" or electrical defects inline has come to the development of inspection systems based on EBI in the past several years. Such technology has developed successfully beyond an engineering analysis tool to the point of allowing cost-effective line monitoring [16,20]. This technology not only provides the ability for higher sensitivity due to smaller pixels and higher resolution imaging but also provides one of the important uses which has been the capability of voltage contrast (VC) imaging, especially for multi-level metal structures. In fact, as brightfield inspection technology capabilities have improved with smaller pixels and UV and DUV illumination the primary application of EBI for line monitoring has been its application to multi-level metal and contact/via levels to detect electrical defectivity using VC

phenomenon. In addition the increased usage of damascene Cu process technology where voids, under-etched vias or other electrical defect mechanisms are increasingly hidden to the optical inspection methods has created a gap in the inline inspection characterization of defectivity that EBI is well-suited for [21]. As a result this technology has seen extensive use for post-Cu CMP (Metal 1 and Metal 2) inspection as well as post-W CMP (contacts) inspection. Examples of defects detected by EBI using VC (both bright and dark) and shown in Figure 27.16 providing both the image from EBI as well as the determination of the cause by focused ion beam (FIB) analysis.

From a detection and inspection standpoint EBI is essentially a brightfield-analogous inspection technology which uses electron beam imaging instead of optical imaging with all the associated issues (for example, charging potential). Analogous to "reflected light" for bright field inspection would be the emitted and detected secondary electrons. Considerations for proper detection optimizing in this case are the beam landing energy, beam current, and field conditions imposed, which create either a retarding or a extracting condition for the emitted electrons. Materials issues play a large role in EBI applications. Insulating materials can lead to charging which inhibits effective inspection and must be countered. Some materials may interact with a given beam energy which could result in defectivity itself. Proper consideration should be given if using such technology on materials such as resist or low-K dielectric layers.

Electron beam inspection has advantages over the traditional electrical probe techniques due to its ability to resolve electrical issues below the typical resolution of probe and the ability to provide accurate defect locations compared to having large area probe structures which require follow-on physical analysis to determine the nature and location of the fail and depend to some degree on engineering judgment as to the cause. Electron beam inspection has been demonstrated to effectively detect defects causing open vias inline that did not cause sufficient signal at inline electrical probe but did result in yield loss at functional test [16].

The HARI structures provide another opportunity for the unique application of EBI since residues or under-etch conditions at the bottom of such structures may respond to e-beam methods while being

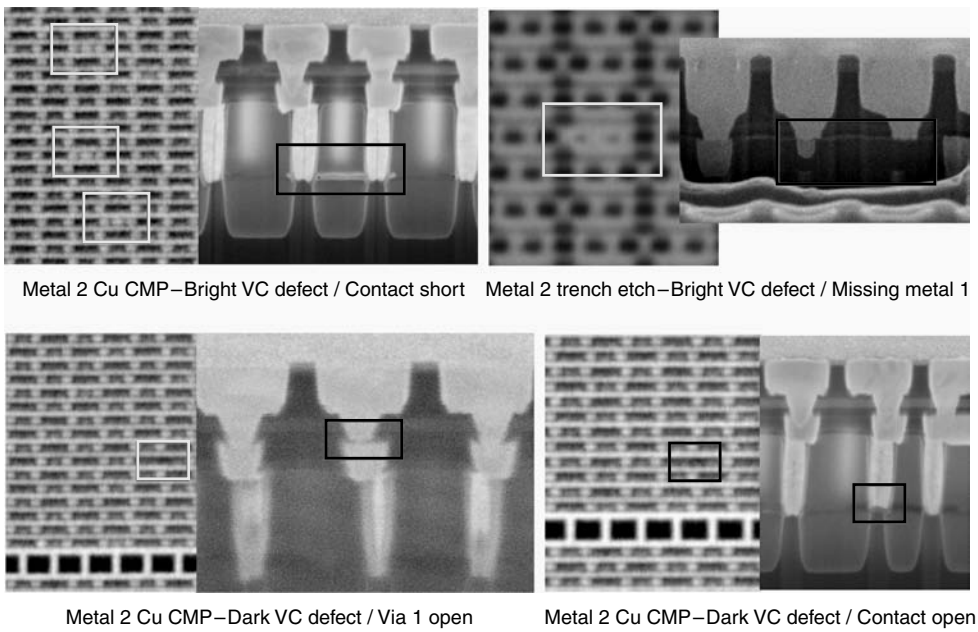

Metal 2 Cu CMP–Bright VC defect / Contact short
Metal 2 trench etch–Bright VC defect / Missing metal 1
Metal 2 Cu CMP–Dark VC defect / Via 1 open
Metal 2 Cu CMP–Dark VC defect / Contact open

FIGURE 27.16 Examples of voltage contrast (VC) defects from electron beam inspection (EBI) and the corresponding results of the focused ion beam (FIB) cross-section.

beyond the detection capability of optical methods. Examples of such structures include contacts, vias, and deep trenches for memory.

Other applications of EBI include inspection of contacts and vias after etch, inspection of gates after gate etch to highlight electrically shorted gates [17], inspection of post-silicidation to find silicide shorts or other silicide-related defectivity and monitoring and reduction of process-induced substrate dislocation defectivity [18]. Since EBI is heavily used in detection of underlying or non-visual electrical defects this technology has increased the need and use of FIB defect analysis. FIB analysis allows one to cross-section defective areas to provide a visual analysis of the cause of the electrical fault. In conjunction with EBI, this analytical technique can provide a significant level of quick feedback for electrical defects that would normally not be detected until final probe [19].

27.3.3.4 Macro Inspection

As the industry moves to more and more automation in the overall process and the post-process verification requires consistency and quantitative analysis and there is a need to remove the human element in these metrology steps. As a result there is a market for a lower sensitivity inspection tool. This tool could effectively remove the historic needs for human-operated microscope evaluations for after develop (lithography) inspection among other applications. A macro inspection tool, as it is commonly referred, operates by taking a low resolution image of the entire wafer and comparing that image to a stored image of an "ideal" wafer. A typical resolution capability of such tools is 30–50 μm.

Some of the application and capabilities of the macro inspection tools are the detection of large defect events, verification of proper reticle (in some cases), checking for out-of-focus conditions from lithography (such as hot spots and reticle field tilt), scratches, and back end of line applications, such as post-polyimide or other final layers to detect large defect issues. Since these tools typically can inspect a wafer very quickly these systems can potentially be applied at most any process to provide a continuous check for process stability and variation post-processing. Also since many current yield issues that affect the fabs are events that occur randomly at low levels whether it would be a few wafers in a lot or on a percentage of lots the ability to monitor a large number of wafers and lots is attractive as long as the system is capable of detecting the problems.

Due to the ability to create such inspection systems with minimal overall size these systems can be incorporated into the process tools themselves or they can be acquired as stand-alone systems, the former being a preferable configuration for unit process monitoring, while the latter allowing more general accessibility. These systems can come close to the ideal of having a real-time process defectivity monitor on product wafers albeit without a high level of sensitivity to many defect issues. A drawback is the need to create inspection recipe setups for every device and each process level in the case of a lithography tool, for instance.

27.3.3.5 Wafer Edge Inspection

Relatively newer to the market are tools to inspect the region of the wafer edge outside of the printed area of the wafer. In order to obtain high, competitive yields in today's industry the yield of the die at the edge of the wafer becomes strategically important. To accommodate this need and the need for consistent and reliable data these tools have been developed. Often the source of defectivity that results in loss of edge die or even inner die is directly attributed to defects that originate from the edge bead exclusion region and even the bevel edge of the wafer. Most work to reduce this defectivity has been done by either drawing a correlation to defectivity seen further in the wafer (and thus using the previously existing inspection tools) or by manually using a microscope or SEM to investigate the wafer edge. Now there are tools to perform automated inspection of the wafer edge and bevel and provide meaningful data on the defectivity seen.

Typical defects of interest are peeling or delaminating films, residual films, chips, cracks, and particles. Like the macro inspection tool the wafer edge inspection tools will attempt to replace a typically manual application with automation that can provide better quantitative results where that is needed. It can also

be applied to a large number of lots and wafers without redirecting manpower resources to this effort during those instances when evaluation of the wafer edge is needed.

27.3.4 Yield Impact Prediction/Verification

Only a percentage of defects observed in line cause device failures because many defects land on non-critical areas of the chip, or are of a size smaller than the minimum geometry, or are of a material that does not cause electrical shorting. "Kill ratio" for a process level is defined as the ratio of the number of killer defects to the number of observed defects at that level. The Kill ratio has to be calculated to estimate yield loss associated with any product inspection. To obtain the kill ratios, electrical test wafer maps are overlaid on in-line defect maps. One-to-one correlation of a defect location on both maps indicates a killer defect. This procedure is easier in concept than in practice sometimes, multiple defects are observed at a fail location, while at other times no physical defects are observed. This correlation of in-line to end-of line probe data were easier with memory chips (or memory areas of a chip) where detailed bit maps can be generated. For more complex logic chips, this correlation is usually done with bin level test data or bit map data from the cache memory area. These days there are yield management software systems that can automatically generate these correlations. In the absence of this correlation, defect classification information from optical or SEM review is the next best way to experimentally obtain kill ratio.

Critical area estimators are also available to model the percent of a chip area that is sensitive to defects of different defect sizes. The most common model uses "dot throwers" where the model simulates imaginary defects dropping randomly over the area of the chip design database. The model then runs these defect distributions through a "shapes-checker" to look for certain fail criteria, such as shorts, blocks, partial opens, etc. These models provide probability of failure curves as a function of defect size for different device levels. Figure 27.17 is an example of a family of these curves for a 0.5 μm device [6]. In line inspection data provide a plot of defect count vs. defect size. The product of these probability curves with inline defect vs. size plots for the same device levels equals the number of "kills" at each defect size.

Martin [7] defined a term "defect density learning rate (LR)" to quantify the rate of yield improvement and excursion level in a factory.

$$\mathrm{LR} = \{1 - (D_\mathrm{n}/D_\mathrm{o})^{1/n}\}100\% \tag{27.3}$$

where n, time period; D_n, new defect density; D_o, old defect density

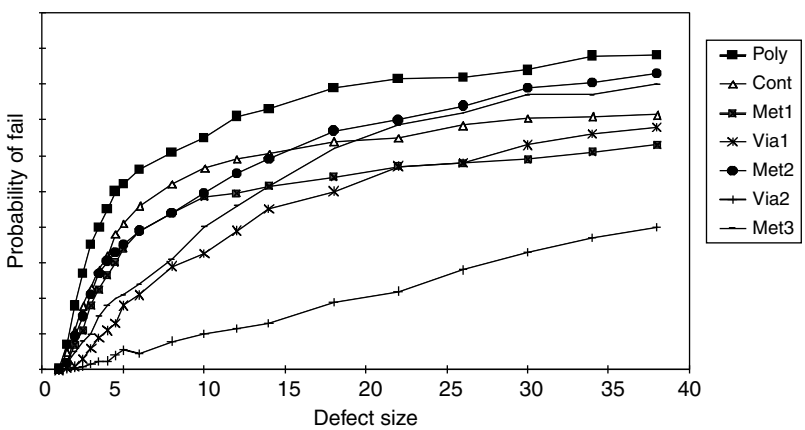


FIGURE 27.17 TLM device of 0.5 μm critical area curves. (From Winter, T., *Proceedings of SPIE*, Vol. 3215, Austin, TX, 1997, 62.)

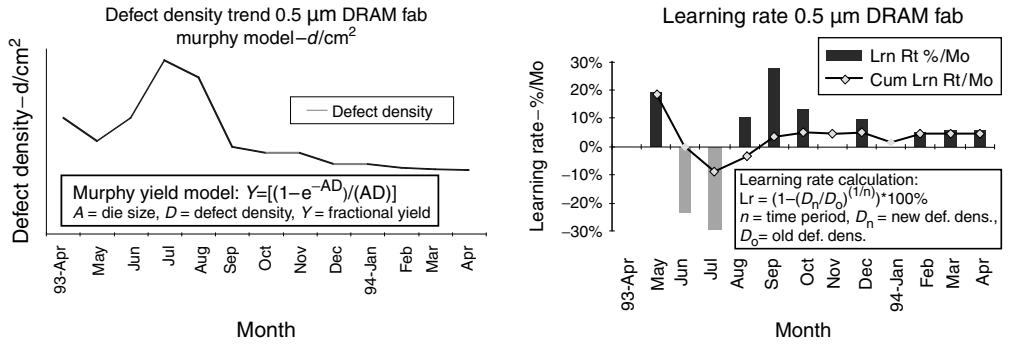


FIGURE 27.18 Measuring a fab's yield methodology and results. (From Martin, R., *KLA Yield Management Seminar*, Geneva, Switzerland, 1996.)

Figure 27.18 shows an example of LR for a 0.5 µm dynamic random access memory (DRAM) fab [7]. Figure 27.18a shows the defect density over a period of time, while Figure 27.18b is a calculation of LR. In this case, the fab had a cumulative LR of 5% and 4 excursions in 15 months or an excursion rate of 27%. The LR is expected to be high during the early portion of new product introduction, but it tapers off as the fab approaches yield entitlement. The defect density at yield entitlement is the "best realistic" defect density capability of a fab for a device, and is dependent on design rule, layout efficiency, die per wafer and equipment set capability. This entitlement defect density is empirically determined from past experience, best wafer/lot yield analysis and benchmarking against other devices or companies. Martin [7] has published empirical curves of entitlement defect density vs. polysilicon design rule for different generations of equipment capability (see Figure 27.19).

In-line to end-of-line defect correlators, critical area extraction, spatial signature analysis routines and defect partitioning software are elements of an overall fab yield management software system. Such a system must not only automatically identify process steps at which critical defects occur, but also it must

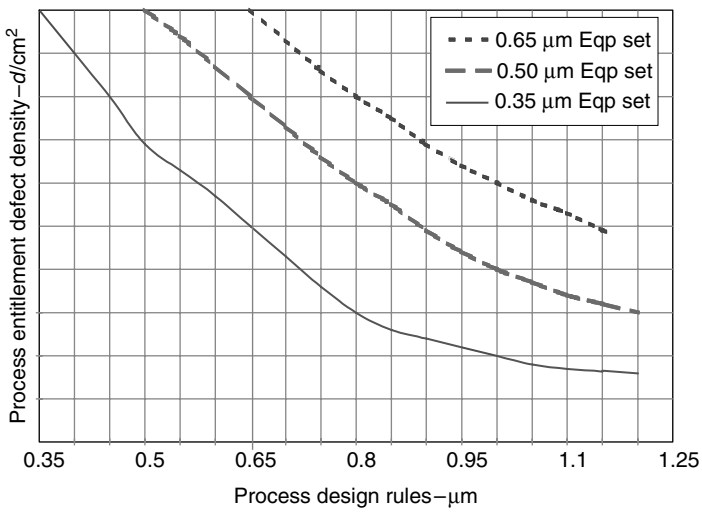


FIGURE 27.19 Empirical curves of process entitlement defect density vs. process design rule. (From Martin, R., *KLA Yield Management Seminar*, Geneva, Switzerland, 1996.)

predict yield impact and couple to defect knowledge databases or decision trees to direct operators to specific corrective actions. Some of these capabilities are currently available, others are imminent because of improved software and networking systems.

27.3.5 Root Cause Isolation

The purpose of root cause isolation is to determine the sources of problematic defects so that the efforts of the process engineers may be focused on specific areas of the process to fix the defect. Root cause isolation is made up of two parts: (1) prioritization of defect types and (2) isolation of the defect source.

The prioritization of defect types requires consistent defect detection and classification, either manual or automatic, to understand the levels of specific defect types. Defect paretos can be created to prioritize defects by level and type. From the paretos, the process level and defect type(s) may be chosen for the defect isolation effort.

The isolation of defect sources can utilize defect information from many sources. Equipment monitoring charts, in situ equipment sensor data and equipment commonality analysis may lead to the errant equipment and process. Sometimes, a library (or database) of commonly observed defect images and energy dispersive x-ray (EDX) spectra can be very useful in the defects. For example, Figure 27.20 shows the EDX spectra of and etch O-ring. However, further detailed analyses using SEM, EDX, FIB, and other analytical tools may be required to track the problem to a component within a specific piece of equipment. A tool for root source isolation is the use of Wafer Positional Analysis [22]. For this analysis to be effective the wafer fab must track the wafer positions in the cassette throughout the process flow and randomly sort the wafers by cassette position at a variety of pre-determined points in the flow. From this database a number of metrics, such as yield, can be plotted by the wafer position to determine if a pattern is visible (e.g., odd–even yield pattern) by the wafer position for a section of the process. This can focus the root cause efforts to a particular section of the process flow with the resolution limited by the number of random sorting steps.

Often, defect isolation requires scanning and reviewing wafers at different steps in a process sector to understand the sources of the defects. This is done by from the following process:

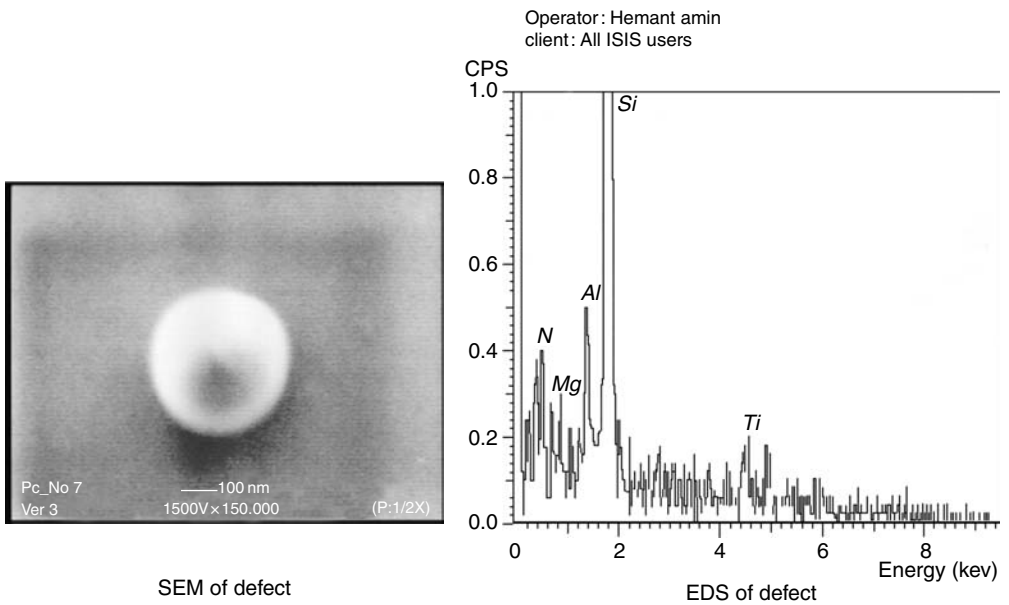


FIGURE 27.20 Scanning electron microscopy (SEM) image and energy dispersive x-ray (EDX) spectrum of an oxide etcher o-ring-induced wafer defect.

1. Scanning and reviewing wafers at a step prior to the first step of the defect source. This captures the incoming defects to the isolation process.
2. Scanning wafers and reviewing the adders at each possible point of the defect source. Scan and review the same wafers inspected at the previous steps.
3. Scanning wafers and reviewing all defects (adders and defects common to the previous steps) at the last step in the isolation procedure. Create a pareto of all defect types. For the chosen defect types, extract the point in the inspection steps where the defects first appeared. This should point to the source of the defect.

Another iteration of this procedure on a narrow range of the process may be necessary to determine the source of the defects to specific tools at specific steps.

This phase of yield enhancement usually takes the longest time since it requires slow and systematic "forensic" analyses. Having identified the problem, implementation of a solution can be fairly quick. The fix may be a design change, a simple component change, a chamber clean or a process recipe modification. Once a solution is found, several tests need to be run to verify the fix and final verification is usually based on in-line defect levels and electrical test results.

27.4 Summary

Rapid yield ramps will continue to require continuous improvement in defect levels from processes and equipment along with a sound yield management methodology. In this chapter, yield management methods were reviewed along with details about defect sources, process/equipment control and the application of defect detection and analysis tools, yield models and software for yield enhancement. Overall strategies for yield management during various stages of new process development were also covered.

References

1. *The International Technology Roadmap for Semiconductors*, Austin, TX: International SEMATECH, 2003 Edition. (See also http://public.itrs.net/, accessed on 9 February 2007).
2. Akella, R., W. Jang, W. W. Kuo, R. K. Nurani, and E. H. Wang. "Defect Sampling Strategies for Yield Management." In *KLA Yield Management Seminar*, Santa Clara, CA, 1996.
3. Williams, R. and R. Nurani. "Sampling Plan Optimization for a Multi-product Scenario in Semiconductor Manufacturing." In *KLA-Tencor Yield Management Seminar*, SEMICON/West, San Francisco, CA, 1997.
4. McIntyre, M., R. Nurani, R. Akella, and A. Strojwas. "Key Considerations in Sampling Methodologies and Yield Prediction." In *KLA Yield Management Seminar*, Makuhari, Japan, 1996.
5. Bennett, M., J. Garvin, J. Hightower, Y. Moalem, and M. Reddy. "The Matching of Multiple IMPACT Systems in Production." In *KLA Yield Management Seminar*, San Francisco, CA, 1997.
6. Winter, T. "Electrical Defect Density Modeling for Different Technology Nodes, Process Complexity and Critical Areas." In *Proceedings of SPIE*, Vol. 3215, 62, Austin, TX, 1997.
7. Martin, R. "Benchmarking Fab Yield Opportunities." In *KLA Yield Management Seminar*, Geneva, Switzerland, 1996.
8. Borden, P. *Microcontamination* 9, no. 1 (1991): 43.
9. Menon, V., and M. Grobelny. "Recent Experiences with In Situ Contamination Monitoring and Control." In *AVS Symposium*, San Jose, CA, 1997.
10. Takahashi, K., and J. Daugherty. "Current Capabilities and Limitations of In Situ Particle Monitors in Silicon Processing Equipment." *J. Vac. Sci. Technol., A* 14, no. 6 (1996): 2983–93.
11. Guldi, R. "In-Line Defect Reduction From a Historical Perspective and Its Implications for Future Integrated Circuit Manufacturing." *IEEE Trans. Semicond. Manuf.* 17, no. 4 (2004): 629–40.

12. Nelson, D., and G. Stark. "How Automated Visual Inspection and CD Metrology Will Impact Wafer-Level Packaging." *Chip Scale Rev.* July (2002): 63–9.
13. Cheema, L. A., L. Olmer, and O. D. Patterson. "Wafer Back Side Inspection Applications for Yield Protection and Enhancement." In *Proceedings of the 2002 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 30 April–5 May, 64–71, Boston, MA, 2002.
14. Lederer, K., M. Scholze, U. Strohbach, A. Wocko, T. Reuter, and A. Schoenauer. "Wafer Backside Inspection Applications in Lithography." In *Proceedings of the 2003 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 31 March–1 April, 1–8, Munich, Germany, 2003.
15. Cheema, L. A., L. Olmer, O. D. Patterson, S. Lopez, and M. Burns. "Yield Enhancement from Wafer Backside Inspection." *Solid State Technol.* 46, no. 9 (2003): 57–60.
16. Soucek, M., J. Anderson, H. Chahal, D. W. Price, K. Boahen, and L. Breaux. "Electrical Line Monitoring in a 300 mm Copper Fab." *Semicond. Int.* 26, no. 8 (2003): 80–90.
17. Patterson, O. D., B. Crevasee, K. Harris, B. B. Patel, and G. Cochran. "Reducing Gate-Level Electrical Defectivity Rapidly Using Voltage-Contrast Test Structures." *Micro Mag.* 21, no. 8 (2003): 45–55.
18. Baltzinger, J.-L., S. Desmercieres, S. Lasserre, P. Champonnois, and M. Mercier. "E-Beam Inspection of Disclocations: Product Monitoring and Process Change Validation." In *Proceedings of the 2004 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 4–6 May, 359–66, Boston, MA, 2004.
19. Rathert, J., and J. Teshima. "Dual Approach to Understanding Failure." *Eur. Semicond.* 24, no. 6 (2002): 41–4.
20. Ache, A., and K. Wu. "Production Implementation of State-of-the-Art Electron Beam Inspection." In *Proceedings of the IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 4–6 May, 344–7. Boston, MA, 2004.
21. Henry, T., D. W. Price, and R. Fiordalice. "E-Beam Inspection: Best Practices for Copper Logic and Foundry Fabs." In *Proceedings of the 2003 IEEE International Symposium on Semiconductor Manufacturing*, 30 September–2 October, 396–9. San Jose, CA, 2003.
22. Kittler, R., M. McIntyre, C. Bode, T. Sonderman, S. Reeves, and S. Zika. "Achieving Rapid Yield Improvement." *Semicond. Int.* 27, no. 8 (2004): 53–60.
23. Jarvis, R., and M. Retersdorf. "Can Technology Keep Pace with High Aspect Ratio Inspection?" *Solid State Technol.* 46, no. 11 (2003): 49–50 see also 52.
24. Goel, H., and D. Dance. "Yield Enhancement Challenges for 90 nm and Beyond." In *Proceedings of the 2003 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 31 March–1 April, 262–5, Munich, Germany, 2003.

28

Electrical, Physical, and Chemical Characterization

Dieter K. Schroder
Arizona State University

Bruno W. Schueler
Revera Inc.

Thomas Shaffner
National Institute of Standards and Technology

Greg S. Strossman
Evans Analytical Group LLC

28.1 Introduction **28**-1
28.2 Electrical Characterization **28**-2
Resistivity and Carrier Concentration • MOSFET Device Characterization • Oxide Charges, Interface States, and Oxide Integrity • Defects and Carrier Lifetimes • Charge-Based Measurements • Probe Measurements
28.3 Physical and Chemical Characterization **28**-30
High Spatial Resolution Imaging • Dopants and Impurities • Surface and Thin Film Composition and Chemistry • Stress and Physical Defects
References **28**-62

28.1 Introduction

Most of the characterization techniques described in this chapter are typically applied outside the semiconductor cleanroom environment, at on- or off-site analytical laboratories. These include a variety of electrical measurements, as well as some of the more widely used physical, chemical, and high-resolution imaging procedures, which are applied routinely to manufacturing in a problem-solving mode. They are typically not routine enough for real-time applications inside the wafer fab, mostly because of technique complexity and difficulty in interpreting the data. However, some newer instruments have been adapted for use in fabs, usually in a *near-line* configuration as opposed to an *in-line* one.

Measurements required at the front end, such as critical dimension (CD) and overlay, film thickness, and wafer contamination, can be considered in-line, where immediate feedback for process refinement is required. Reviews of these tools can be found in Chapter 24 and Chapter 25 of this handbook. Likewise, the diagnostics of field returns in a failure analysis mode involve yet another class of analysis and tools, which is discussed in Chapter 29.

The collection of characterization instruments for near-line or analytical laboratory applications continues to grow in number and specialization, and it is therefore necessary to limit the scope of this chapter to only those most frequently used. A number of good books and encyclopedias are available for further study [1–4]. Also, characterization symposium proceedings are routinely published by the Materials Research Society, the American Vacuum Society, the Electrochemical Society, the American Physical Society, the American Chemical Society, and the Society for Photo-Optical Instrumentation Engineers. Other international conferences specialize in ion mass spectrometry [5], electron microscopy

[6], x-ray diffraction (XRD) [7], scanning probe microscopy (SPM) [8], and characterization overviews aligned to the National Technology Roadmap for Semiconductors [9].

The format adopted in this chapter is designed to help the reader quickly extract the basic purpose and concepts fundamental to the operation of each technique. This is followed by a brief description of the most prominent strengths as well as shortcomings. The respective headings, *Purpose, Method, Strengths, and Weaknesses* apply throughout the chapter. Key references are provided to guide readers who are interested in a more detailed study.

28.2 Electrical Characterization

28.2.1 Resistivity and Carrier Concentration

28.2.1.1 Four-Point Probe

Purpose. The *four-point probe* technique is the most common method for measuring the semiconductor resistivity and sheet resistance [3]. It is an absolute measurement without recourse to calibrated standards and is most commonly used to generate sheet resistance contour maps.

Method. Two-point probe methods would appear to be easier to implement, because only two probes need to be manipulated, but the interpretation of the measured data is more difficult. For the two-point probe or two-contact arrangement of Figure 28.1a, each contact serves as a current *and* as a voltage probe. The total resistance R_T is given by

$$R_T = V/I = 2R_W + 2R_C + R_{DUT} \tag{28.1}$$

where R_W is the wire and probe resistance, R_C the contact resistance, and R_{DUT} the resistance of the device under test. Clearly, it is impossible to determine R_{DUT} with this measurement arrangement. Although the current path in the four-contact arrangement in Figure 28.1b is identical to that in Figure 28.1a, the voltage is now measured with two additional contacts. Although the voltage path contains R_W and R_C as well, the current flowing through the voltage path is very low due to the high input impedance of the voltmeter (around $10^{12}\ \Omega$ or higher), making the voltage drops across R_W and R_C negligibly small, and the measured voltage is essentially the voltage drop across the DUT. Using four, rather than two, probes, we have eliminated parasitic voltage drops. Such four contact measurements are usually referred to as *Kelvin measurements*, after Lord Kelvin.

Typical probe radii are 30–500 μm and probe spacing ranges from 0.5 to 1.5 mm. Smaller probe spacing allows measurements closer to wafer edges, an important consideration during wafer mapping. Probes to measure metal films should not be mixed with probes to measure semiconductors. For some applications, e.g., magnetic tunnel junctions, polymer films, and semiconductor defects, microscopic

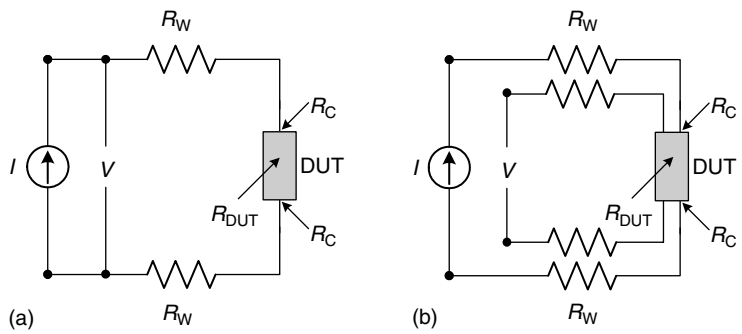


FIGURE 28.1 Two-terminal and four-terminal resistance measurement arrangements.

four-point probes with 1.5 μm probe spacing have been used [10]. A four-point probe consisting of independently driven actuators for use in a scanning electron microscope (SEM) with probe spacing from 18 to 500 μm has been developed [11].

For an arbitrarily shaped sample, the resistivity is given by

$$\rho = 2\pi s F(V/I)\ [\text{ohm} - \text{cm}] \tag{28.2}$$

where s is the probe spacing and F is a correction factor that depends on the sample geometry. F corrects for probe location near sample edges, for sample thickness, sample diameter, probe placement, and sample temperature. For *collinear* or *in-line probes* with equal probe spacing, the wafer thickness correction factor F is [12]

$$F = \frac{t/s}{2\ln\{[\sin h(t/s)]/[\sin h(t/2s)]\}} \tag{28.3}$$

for a *non-conducting* bottom wafer surface boundary, where t is the wafer or layer thickness. For thin, uniformly doped samples, $t \leq s/2$, the resistivity ρ and the sheet resistance R_{sh} are given as

$$\rho = \frac{\pi}{\ln 2} t \frac{V}{I} = 4.532 t \frac{V}{I}\ [\text{ohm} - \text{cm}]; \qquad R_{\text{sh}} = \frac{\rho}{t} = \frac{\pi}{\ln 2}\frac{V}{I} = 4.532\frac{V}{I}\ [\text{ohms/square}] \tag{28.4}$$

The sheet resistance for non-uniform samples of thickness t is

$$R_{\text{sh}} = \frac{1}{\int_0^t [1/\rho(x)]\mathrm{d}x} = \frac{1}{\int_0^t \sigma(x)\mathrm{d}x} = \frac{1}{q\int_0^t [n(x)\mu_n(x) + p(x)\mu_p(x)]\mathrm{d}x} \tag{28.5}$$

where σ is the conductivity. R_{sh} is proportional to the *integrated* conductivity or implant dose. A sheet resistance measurement integrates the entire doping density profile into one simple measurement. An example of contour maps is shown in Figure 28.2.

The key to high-precision four-point probe measurements is the use of two measurement configurations at each probe location, known as the "dual configuration" or the "configuration switched" method [13]. The first configuration is usually with current into probe 1 and out of probe 4 and voltage sensed across probes 2 and 3. The second measurement is made with current driven through probes 1

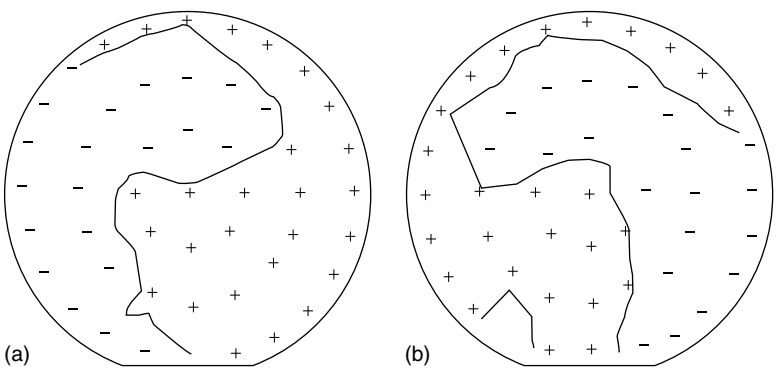


FIGURE 28.2 Four-point probe sheet resistance contour maps; (a) boron, 10^{15} cm^{-2}, 40 keV, $R_{\text{sh,av}} = 98.5\ \Omega$ per square; (b) arsenic, 10^{15} cm^{-2}, 80 keV, $R_{\text{sh,av}} = 98.7\ \Omega$ per square; 1% intervals. 200 mm diameter Si wafers. Data courtesy of Marylou Meloni, Varian Ion Implant Systems.

and 3 and voltage measured across probes 2 and 4. The advantages are that the probes no longer need to be in a high-symmetry orientation (being perpendicular or parallel to the wafer radius of a circular wafer or to the length or width of a rectangular sample), the lateral dimensions of the specimen do not have to be known since the geometric correction factor results directly from the two measurements, and the two measurements self-correct for the actual probe spacing.

Although the collinear probe configuration is the most common four-point probe arrangement, other arrangements are also possible. Arrangement of the points in a square has the advantage of a smaller area. The square arrangement is more commonly used, not as an array of four mechanical probes, but rather as contacts to square semiconductor samples. The theoretical foundation of measurements on irregularly shaped samples is based on conformal mapping developed by van der Pauw [14], provided the contacts are at the circumference of the sample, the contacts are sufficiently small, the sample is uniformly thick, and the surface of the sample is singly connected, i.e., the sample does not contain any isolated holes.

For the flat sample of a conducting material of arbitrary shape, with contacts 1, 2, 3, and 4 along the periphery, as shown in Figure 28.3, the resistance $R_{12,34}$ is defined as

$$R_{12,34} = \frac{V_{34}}{I_{12}} \tag{28.6}$$

where the current I_{12} enters the sample through contact 1 and leaves through contact 2 and $V_{34} = V_3 - V_4$ is the voltage difference between the contacts 3 and 4. $R_{23,41}$ is defined similarly.

The resistivity is given by

$$\rho = \frac{\pi}{\ln(2)} t \frac{(R_{12,34} + R_{23,41})}{2} F \tag{28.7}$$

where F is a function only of the ratio $R_r = R_{12,34}/R_{23,41}$, satisfying the relation

$$\frac{R_r - 1}{R_r + 1} = \frac{F}{\ln(2)} arcosh\left(\frac{\exp[\ln(2)/F]}{2}\right) \tag{28.8}$$

For symmetrical samples such as the circle or square in Figure 28.4, $R_r = 1$ and $F = 1$.

The van der Pauw equations are based on the assumption of negligibly small contacts located on the sample periphery. Real contacts have finite dimensions and may not be exactly on the periphery of the sample. Corner contacts introduce less error than contacts placed in the center of the sample sides. However, if the contact length is less than about 10% of the side length, the correction is negligible for either contact placement [15].

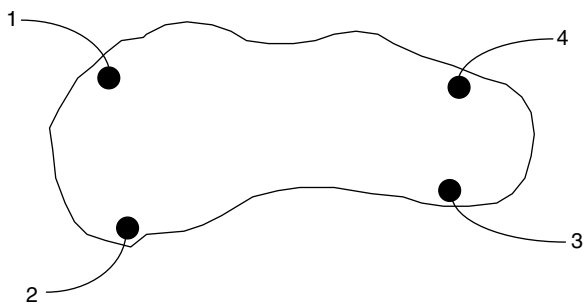


FIGURE 28.3 Arbitrarily shaped sample with four contacts.

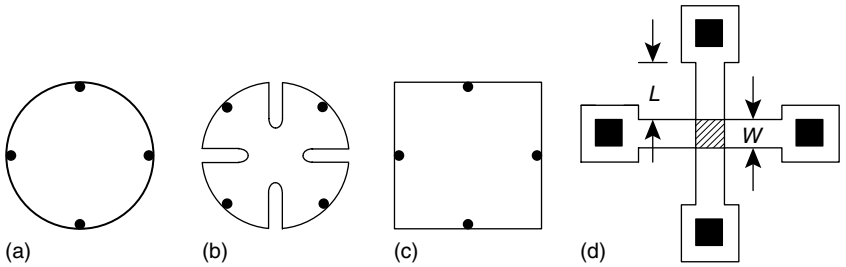


FIGURE 28.4 Typical symmetrical circular and square sample geometries.

Geometries other than those in Figure 28.4a and b are also used. One of these is the *Greek cross* in Figure 28.4c. Using photolithographic techniques, it is possible to make such structures very small and place many of them on a wafer for uniformity characterization. The sheet resistance of the shaded area is determined in such measurements. For structures with $L = W$, the contacts should be placed so that $d \leq L/6$ from the edge of the cross, where d is the distance of the contact from the edge [16]. Surface leakage can introduce errors if L is too large [17]. A variety of cross-sheet resistor structures have been investigated and their performance compares well with conventional bridge-type structures [18]. The measured voltages in cross and van der Pauw structures are lower than those in conventional bridge structures.

The cross and the bridge structures are combined in the cross—bridge structure in Figure 28.5, allowing the *sheet resistance* and the *linewidth* to be determined [19]. Such measurements have shown high levels of repeatability. For a linewidth of 1 μm, the repeatability has been demonstrated to be on the order of 1 nm [20]. Precisions of 0.005 μm and lines as narrow as 0.1 μm have been measured. The sheet resistance is given by

$$R_{\text{sh}} = \frac{\pi}{\ln 2} \frac{V_{34}}{I_{12}} \tag{28.9}$$

where $V_{34} = V_3 - V_4$ and I_{12} is the current flowing into contact I_1 and out of contact I_2.

The right part of Figure 28.5 is a bridge resistor to determine the linewidth W. The voltage along the bridge resistor is

$$V_{45} = \frac{R_{\text{sh}} L I_{26}}{W} \tag{28.10}$$

where $V_{45} = V_4 - V_5$ and I_{26} is the current flowing from contact 2 to contact 6. From Equation 28.10, the linewidth is

$$W = \frac{R_{\text{sh}} L I_{26}}{V_{45}} \tag{28.11}$$

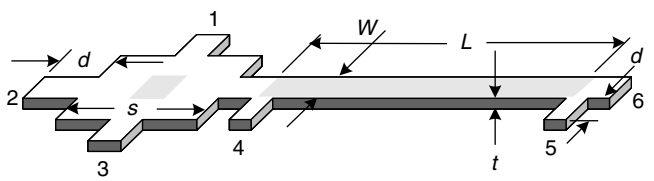


FIGURE 28.5 A cross-bridge sheet resistance and linewidth test structure.

with R_{sh} determined from the cross structure and Equation 28.9. A key assumption in this measurement is that the sheet resistance be identical for the entire test structure.

Since the bridge structure in Figure 28.5 is suitable for resistance measurements, it can be used to characterize "dishing" during chemical–mechanical polishing of semiconductor wafers, where soft metal lines tend to polish thinner in the central portion than at the edges leading to non-uniform thickness. This is particularly important for soft metals such as copper. With the resistance inversely proportional to metal thickness, resistance measurements can be used to determine the amount of dishing [21].

An assumption in Equation 28.11 is that the sheet resistance in the bridge portion of the test structure is the same as that in the cross portion, i.e., in both cross-hatched areas. If that is not true, W will be in error [22]. What exactly is L? Is it the center-to-center spacing as illustrated in Figure 28.5? That depends on the exact layout of the structure. With arms 4 and 5 extending only below the measured line as in Figure 28.5, L is approximately as shown. For symmetrical structures, i.e., arms 4 and 5 extending above as well as below the line, an effective length is $L_{eff} \approx L - W_1$, where W_1 is the arm width. For long structures, i.e., $L \approx 20W$, this correction is negligible, but for short lines, it must be considered, because the contact arms distort the current path. Other considerations are $t \leq W$, $W \leq 0.005L$, $d \geq 2t$, $t \leq 0.03s$, and $s \leq d$ [23].

Strengths and Weaknesses. The method's strength lies in its established use and the fact that it is an absolute measurement without recourse to calibrated standards. It has been used for many years in the semiconductor industry and is well understood. With the advent of wafer mapping, the four-point probe has become a very powerful process-monitoring tool. The equipment is commercially available. The weakness of the four-point probe technique is the surface damage it produces and the metal it deposits on the sample. The damage is not very severe but sufficient not to allow measurements on product wafers. The probe also samples a relatively large volume of the wafer, preventing high-resolution measurements.

28.2.1.2 Modulated Photoreflectance

Purpose. Modulated photoreflectance generates dose contour maps of ion-implanted samples without the need for implant activation, because the contour map is measured immediately after implantation.

Method. It measures the modulation of the optical reflectance of a sample in response to waves that are generated, when a semiconductor sample is subjected to periodic heat stimuli. In the *modulated photoreflectance* or *thermal wave* method, an Ar^+ ion laser beam is modulated at a frequency of 0.1–10 MHz. A periodic temperature variation is established in the semiconductor in response to this periodic heat stimulus, with an amplitude around 10°C in silicon. The thermal wave diffusion length at a 1 MHz modulation frequency is 2–3 μm [24]. The small temperature variations cause small volume changes of the wafer near the surface. These changes include both thermoelastic and optical effects, and they are detected with a second laser—the probe beam—by measuring a reflectivity change. The apparatus and a contour map are illustrated in Figure 28.6. Both pump and probe laser beams are focused to approximately 1 μm diameter spots, giving the technique a spatial resolution of around 1 μm, allowing measurements on patterned wafers.

Modulated photoreflectance is a comparative technique. To convert from thermal wave signal to implant dose requires calibrated standards with known implant doses. The ability to determine ion implant densities by thermal waves depends on the conversion of the single-crystal substrate to a partially disordered layer by the implant process. The thermal-wave-induced thermoelastic and optical effects are changed in proportion to the number of implanted ions. Modulated photoreflectance implant monitoring is subject to post-implant damage relaxation.

The technique is contactless and non-destructive and has been used to measure implant doses from 10^{11} to $10^{15}\,cm^{-2}$ [25]. Measurements can be made on bare and on oxidized wafers. The ability to characterize oxidized samples has the advantage of allowing measurements of implants through an oxide without removing it. The technique can discriminate between implant species, since the lattice damage increases with implant atom size and the thermal wave signal depends on the lattice damage. It has been used for ion implantation monitoring, wafer polish damage, and reactive and plasma etch damage studies.

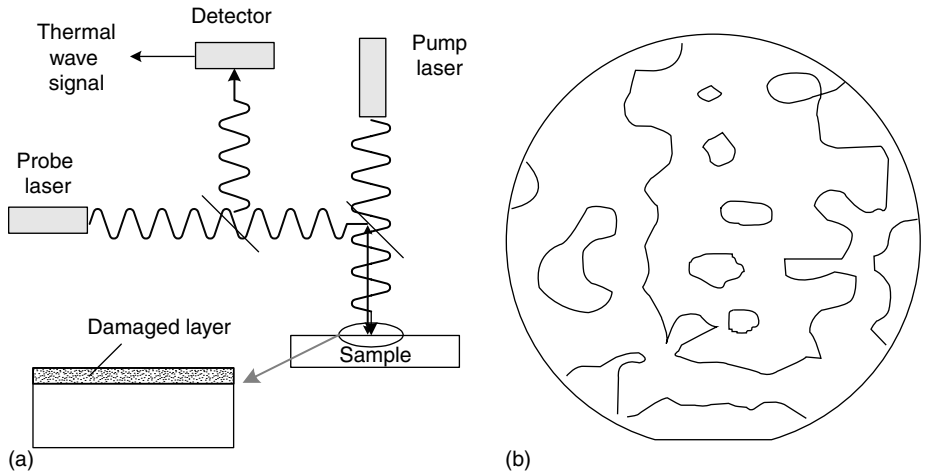


FIGURE 28.6 (a) Schematic diagram of the modulated photoreflectance apparatus and modulated photoreflectance contour maps; (b) boron, 5×10^{12} cm^{-2}, 30 keV, 600 TW units; 0.5% intervals. 200 mm diameter Si wafers. Data courtesy of Marylou Meloni, Varian Ion Implant Systems.

Strengths and Weaknesses. The major strength of modulated photoreflectance is the ability to measure the implant dose immediately after implantation, without wafer annealing. This is a significant time-saver. It has the ability to detect low-dose implants and to display the information as contour maps. The equipment is commercially available. Its main weakness is the qualitative nature of the measurement. The thermal wave signal is proportional to the damage in the sample, but the precise dose is difficult to determine quantitatively.

28.2.1.3 Capacitance–Voltage Profiling

Purpose. The capacitance–voltage (C–V) technique is used to determine the doping density profile of lightly and moderately doped regions.

Method. In the C–V technique, the width of a reverse-biased space-charge region (scr) of a semiconductor junction device [Schottky diode, pn junction, metal-oxide semiconductor (MOS) device] is changed with an applied dc voltage [26]. The capacitance is determined by superimposing a small-amplitude ac voltage on the dc voltage. The ac voltage typically varies at frequencies of 10 kHz to 1 MHz with amplitude of 10–20 mV, but other frequencies and other voltages can be used.

The capacitance of a reverse-biased junction is

$$C = \frac{K_s \varepsilon_0 A}{W} \tag{28.12}$$

where K_s is the semiconductor dielectric constant, ε_0 the permittivity of free space, A the area, and W the scr width. The doping density is related to the capacitance C and dC/dV through the relation

$$N_A(W) = -\frac{C^3}{qK_s\varepsilon_0 A^2 \mathrm{d}C/\mathrm{d}V} = \frac{2}{qK_s\varepsilon_0 A^2 \mathrm{d}(1/C^2)/\mathrm{d}V} \tag{28.13}$$

The region that is profiled is the edge of the reverse-biased scr width, W, given by

$$W = \frac{K_s \varepsilon_0 A}{C} \tag{28.14}$$

The doping density is obtained from a *C–V* curve by taking the slopes d*C*/d*V* or d(1/C^2)/d*V*. For a Schottky barrier diode, there is no ambiguity in the scr width since it can only spread into the substrate. The scr spreading into the metal is totally negligible. The doping profile equations are equally well applicable for asymmetrical p^+n and n^+p junctions.

Metal-oxide semiconductor capacitors (MOS-C) and MOS field effect transisters (MOSFETs) can also be profiled [27] However, care must be taken to eliminate minority carriers. For MOSFETs, minority carriers are eliminated by reverse biasing the source–drain junctions. For a MOS-C, the device must remain in deep depletion during the measurement to eliminate minority carrier contribution to the measurement, ensured with a rapidly varying dc ramp voltage or a pulsed gate voltage. The capacitance is measured immediately after the pulse before minority carriers can be generated. Equation 28.13 applies to MOS-Cs when both interface states and minority carriers can be neglected, but the scr width expression must be modified to

$$W = K_s \varepsilon_0 A \left(\frac{1}{C} - \frac{1}{C_{ox}} \right) \tag{28.15}$$

The spatial resolution of the measured profile is limited by the Debye length, because the capacitance is determined by the movement of majority carriers and the majority carrier distribution cannot follow abrupt spatial changes in dopant density profiles. Detailed calculations show that if a doping density step occurs within one Debye length, the majority carrier density differs appreciably from the doping density profile [28]. The profile limits are determined by Debye length, L_D, considerations near the surface, and by breakdown scr width limits deeper within the sample, as illustrated in Figure 28.7.

A contactless capacitance and doping profiling version uses a contact held in close proximity to the semiconductor wafer. The sensor electrode, 1 mm diameter and coated with high dielectric strength thin film, is surrounded by an independently biased guard electrode. The sensor electrode is held above the wafer by a porous ceramic air bearing, which provides for a very stable distance from the wafer as long as the load on the air bearing does not change, as shown in Figure 28.8. The controlled load is provided by pressurizing a bellows. As air escapes through the porous surface, a cushion of air forms on the wafer, which acts like a spring and prevents the porous surface from touching the wafer. The porosity and air pressure are designed such that the disk floats approximately 0.5 μm above the wafer surface. A stainless steel bellows acts to constrain the pressurized air and to raise the porous disk when the air pressure is

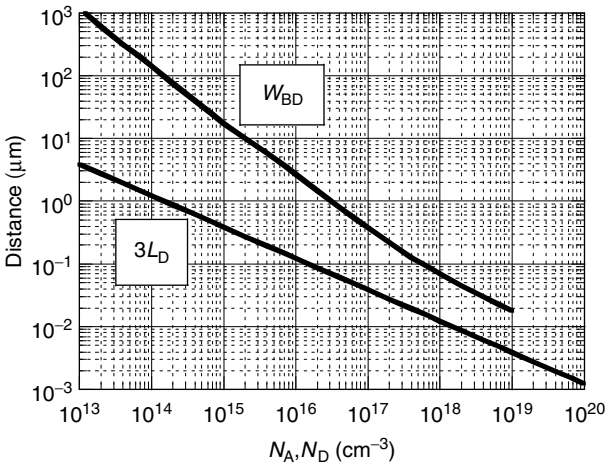


FIGURE 28.7 Spatial profiling limits. The "$3L_D$" line is the lower limit for metal-oxide semiconductor capacitor profiling, and the "W_{BD}" line is the upper profile limit governed by bulk breakdown.

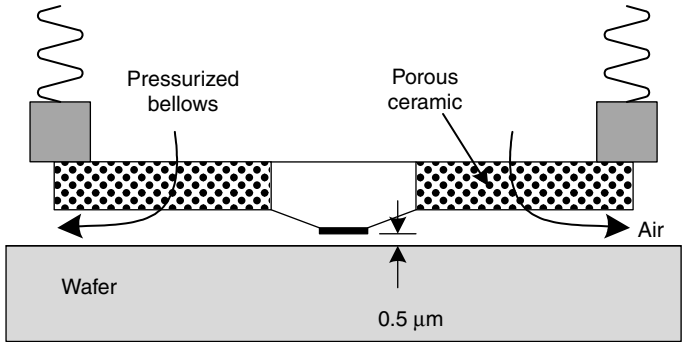


FIGURE 28.8 Contactless doping profiling arrangement. Pressurized air maintains the electrode at approximately 0.5 μm above the sample surface.

reduced. If the air pressure fails, the disk moves up, rather than falling down and damaging the wafer [29].

To prepare the wafer, it is placed in a low-concentration ozone environment at a temperature of about 450°C, reducing the surface charge on the wafer, especially critical for *n*-Si, makes it more uniform, reduces the surface generation velocity, and allows deeper depletion [30]. A recent comparison of epitaxial resistivity profiles by the contactless with Hg-probe *C*–*V* measurements compared very favorably [31]. The capacitance of the air gap is measured by biasing the semiconductor surface in accumulation. Light is used to collapse any possible scr due to surface charge while the sensor is lowered and while the air gap modulation due to the electrostatic attraction is determined to eliminate any series space-charge capacitance. Assuming that the air gap does not vary with changing electrode voltage, the capacitance of the air gap is the measured capacitance at its maximum value. The doping density profile is determined from Equation 28.13 and Equation 28.15 with C_{ox} in Equation 28.15 replaced by C_{air}.

28.2.1.4 Lateral Doping Profiling

The two main techniques that have emerged for *lateral* doping density profiling are *scanning capacitance microscopy* (SCM) and *scanning spreading resistance microscopy* (SSRM) [32]. The SCM has received much attention as a lateral profiling tool [33]. A small-area capacitive probe measures the capacitance of a metal/semiconductor or a MOS contact. The SCM combines atomic force microscopy (AFM) with highly sensitive capacitance measurements. The SCM is able to measure the local *C*–*V* characteristics between the SCM tip and a semiconductor with nanometer resolution. The metallized AFM tip is used for imaging the wafer topography in conventional contact mode and also serves as an electrode for simultaneously measuring the MOS capacitance. The SCM images of actively biased cross-sectional MOSFETs and of operating *pn* junctions allow visualization of the operation of semiconductor devices.

The semiconductor device is usually cleaved or polished so that the device cross-section is exposed, as shown in Figure 28.9, although the sample top, without cleaving, can also be measured. An oxide is deposited on the cross-sectional area and the probe is scanned across the area in the contact mode, measuring the capacitance variations in the nanometric probe/oxide/silicon MOS-C by applying a high-frequency (hf) ac voltage between the probe and the semiconductor. For constant electrical bias, the scr in the MOS-C is wider for lower doping densities. Dedicated simulation models are necessary to obtain a realistic conversion curve that relates the local SCM signal with the local carrier density. A schematic of the measurement in Figure 28.10 shows the conducting AFM tip on the oxidized sample, the *C*–*V*, and d*C*/d*V* curves. The voltage is applied to the substrate in this case. In some cases, it is applied to the tip. The shape of the d*C*/d*V* curve identifies the doping type. The SCM is sensitive to carrier densities from 10^{15} to 10^{20} cm^{-3}, with a lateral resolution of 20–150 nm, depending on tip geometry and dopant density. Extraction of absolute dopant densities requires reverse simulation incorporating tip geometry

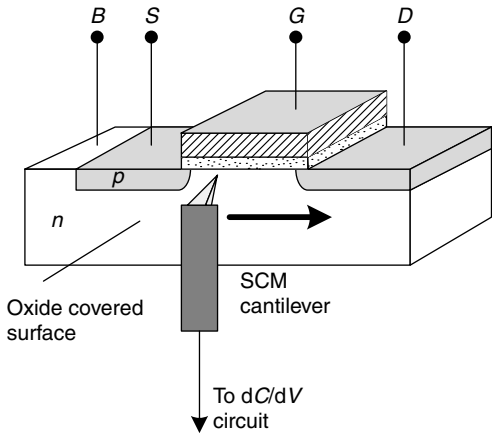


FIGURE 28.9 Scanning capacitance schematic.

and sample oxide thickness. Example SCM maps are given in Figure 28.11, showing the formation of a channel in a MOSFET with increasing gate voltage [34].

Two standard SCM methods have been developed for two-dimensional dopant profiling: in the ΔC mode, a constant amplitude ac bias voltage is applied between the tip and sample, and in the ΔV mode, a feedback loop adjusts the applied ac bias voltage to keep the change in capacitance, ΔC, constant as the tip is moved from one region to another [35]. In the former, the ac bias voltage produces a corresponding change in capacitance measured by a lock-in amplifier. As the tip moves from a region of high dopant density to a more lightly doped region, the lock-in amplifier output increases owing to the larger C–V curve slope in the lightly doped region. In the latter, a feedback loop adjusts the applied ac bias voltage to keep ΔC constant as the tip is moved from one region to another. In this case, the magnitude of the required ac bias voltage is measured to determine the dopant density.

The advantage of the ΔC mode is simplicity. The disadvantage of this system is that a large ac bias voltage (several volts ac) is needed to measure finite SCM signal at high doping densities. When this same

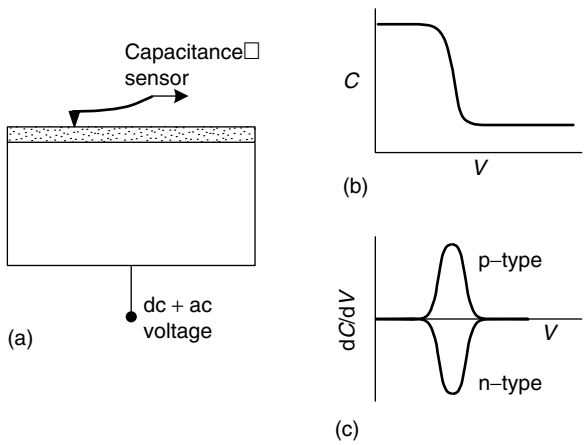


FIGURE 28.10 (a) Schematic of the atomic force microscopy/scanning capacitance microscopy (AFM/SCM) design, (b) C–V curve of n-type substrate with bias applied to the substrate (c) dC/dV curve. The sign identifies the dopant type.

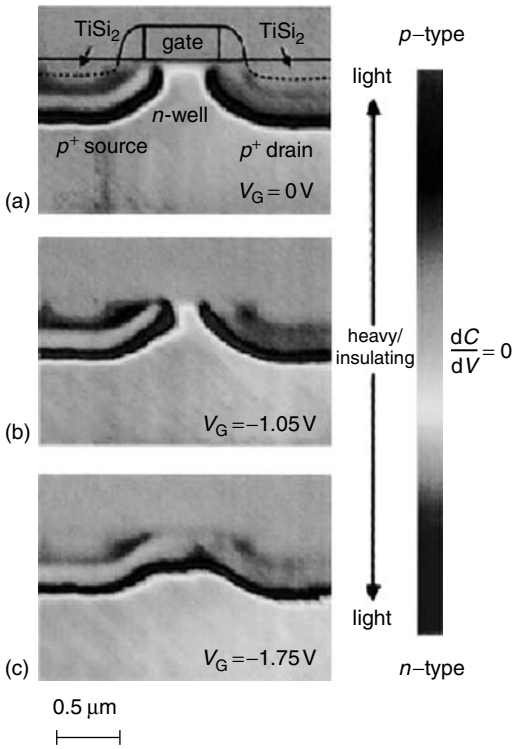


FIGURE 28.11 Sequence of SCM images of an Si, *p*-channel MOSFET with $V_D = -0.1$ V, $V_S = V_B = 0$ V, and $V_G =$ (a) 0, (b) -1.05, and (c) 1.75 V. The progression of the SCM images shows the formation of a conducting channel between the source and the drain. The schematic drawing in (a) shows the approximate locations of the polysilicon gate, titanium nitride spacers, and the titanium silicide contacts. Images were acquired with $V_{ac} = 2.0$ V peak to peak and $V_{dc} = 0$ applied to the SCM tip. (After Nakakura, C.Y., Tangyunyong, P., Hetherington, D.L., and Shaneyfelt, M.R., *Rev. Sci. Instrum.*, 74, 127–133, 2003.)

voltage is applied to lightly doped silicon, it creates a larger depletion volume, reducing the spatial resolution and making accurate modeling more difficult. The advantage of the ΔV method is that the physical geometry of the depletion problem remains relatively constant as the tip is scanned from a lightly to heavily doped region. The disadvantage is that an additional feedback loop is required.

For reproducible measurements, samples must be prepared carefully. Factors that influence the repeatability and the reproducibility of SCM measurements include sample-related problems (mobile and fixed oxide charges, interface states, non-uniform oxide thickness, surface humidity and contamination, sample aging, and water-related oxide traps), tip-related problems (increase of the tip radius, fracture of the tip apex, mechanical wear out of the metal coating, and contaminants on the tip picked up from the sample), and problems related to the electrical operating conditions (amplitude of the ac probing signal in the capacitance sensor, scanning rate, compensation of the stray capacitance, electric-field-induced oxide growth, and dc tip bias voltage).

The SSRM, based on the AFM, uses a small conductive tip to measure the local spreading resistance [36]. The resistance is measured between a sharp conductive tip and a large back surface contact. A precisely controlled force is used while the tip is stepped across the sample. The SSRM sensitivity and dynamic range are similar to conventional spreading resistance. The small contact size and small stepping distance allow measurements on the device cross-section with no probe conditioning. The high spatial

resolution allows direct two-dimensional nanometer spreading resistance profiling (nano-SRP) measurements, without the need for special test structures. Spatial resolution of 3 nm has been demonstrated [37].

For one- or two-dimensional carrier density profile measurements, the sample is cleaved to obtain a cross-section. The cleavage plane is polished using decreasing grit-size abrasive paper and finally colloidal silica to obtain a flat silicon surface. After polishing, the sample is cleaned to eliminate contaminants and finally rinsed in deionized water. The sole limitation is the requirement that the structure be sufficiently wide for the profile in the direction perpendicular to the cross-section of the sample to be uniform.

The AFM equipment is a standard commercially available equipment. A conductive cantilever with a highly doped ion-implanted diamond tip can be used as a resistance probe. Diamond protects the tip from deformation due to the rather high loads ($\sim$50–100 μN) required to penetrate the native oxide layer and make good electrical contact. Coating the tip with a thin tungsten layer improves the conductivity. Like conventional SRP, nano-SRP needs a calibration curve to convert the measured resistances into carrier densities. The resistance is measured at a bias of $\sim$5 mV, as in conventional SRP. Scanning the tip over the cross-section of the sample provides a two-dimensional map of the local spreading resistance with a spatial resolution set by the tip radius of typically 10–15 nm. A straight conversion of spreading resistance to local resistivity is made.

As in conventional spreading resistance measurements, a proper model must be used to interpret the experimental data. It is frequently assumed that the contact between the probe and the sample is ohmic. However, it has been shown that the contact is not ohmic [38]. The *I–V* curves vary from an ohmic-like shape in heavily doped areas to a rectifying in lightly doped areas and that surface states induced by the sample preparation influence the *I–V* curves. The presence of surface states due to sample polishing reduces the current, particularly pronounced in lightly doped areas.

Strengths and Weaknesses. The method's strength lies in its ability to give the carrier density profile with little data processing. A simple differentiation of the *C–V* data suffices. It is an ideal method for moderately doped materials and is non-destructive when a mercury probe is used. It is well established with available commercial equipment. The major weakness of the differential capacitance profiling method is its limited profile depth, limited at the surface by the Debye length and in depth by junction breakdown. The latter limitation is particularly serious for the heavily doped regions. Further limitations are due to the Debye limit for abrupt profiles, which applies to all carrier profiling techniques.

28.2.2 MOSFET Device Characterization

28.2.2.1 Threshold Voltage

Purpose. The measurement of threshold voltage, V_T, is required for process control and for channel length/width and series resistance determination. The threshold voltage for *n*-channel devices, accounting for short- and narrow-channel effects, and ion implantation, is given by

$$V_T = V_{FB} + 2\phi_F + \left(1 - \frac{\alpha}{L}\right) \frac{\sqrt{2qK_s\varepsilon_0 N_A(2\phi_F - V_{BS})}}{C_{ox}} + \frac{qD_i}{C_{ox}} + \frac{\beta}{W} \tag{28.16}$$

where L is the gate length, W the gate width, V_{FB} the flatband voltage, α the short-channel parameter, β the narrow-channel parameter, V_{BS} the substrate–source voltage, and D_i the acceptor implant dose.

Method. One of the most common threshold voltage measurement technique is the *linear extrapolation* method with the drain current measured as a function of gate voltage at a low drain voltage of typically 50–100 mV to ensure operation in the linear MOSFET region [39]. The drain current vs. gate voltage curve is extrapolated to $I_D = 0$ and the threshold voltage is determined from the extrapolated gate voltage V_{Gi} by

$$V_T = V_{Gi} - V_D/2 \tag{28.17}$$

Equation 28.17 is strictly only valid for negligible series resistance. Fortunately, series resistance is usually

negligible at the low drain currents where threshold voltage measurements are made, but it can be appreciable in lightly doped drain (LDD) devices.

The I_D–V_G curve is non-linear at gate voltages below V_T due to subthreshold current and above V_T due to series resistance and mobility degradation effects. It is a common practice to find the point of maximum slope on the I_D–V_G curve by a maximum in the transconductance, $g_m = \Delta I_D/\Delta V_G$, fit a straight line to the I_D–V_G curve at that point and extrapolate to $I_D = 0$, as illustrated in Figure 28.12a giving $V_T = 0.9$ V. Failure to correct for series resistance and mobility degradation leads to an underestimate in V_T [40].

It is obvious from Figure 28.12a that the drain current at the threshold voltage is higher than zero. This is utilized in the *constant drain current* method, where the gate voltage at a specified threshold drain current, I_T, is taken to be the threshold voltage. This technique lends itself readily to threshold voltage mapping. In order to make I_T independent of device geometry, $I_T = I_D/(W_{eff}/L_{eff})$ is sometimes specified at a current around 10–50 nA but other values have been used [41]. In Figure 28.12b, the threshold voltages for $I_D = 1$ μA and $I_D = 10$ μA are shown. It is quite obvious that these two V_Ts differ from each other and from the linear extrapolated value. Nevertheless, the method has found wide application, provided a consistent drain current is chosen.

In the *subthreshold* method, the drain current is measured as a function of gate voltage below threshold and plotted as $\log(I_D)$ vs. V_G. The subthreshold current depends linearly on gate voltage in such a semilog plot. The gate voltage at which the plot departs from linearity is sometimes taken as the threshold voltage. However, for the data of Figure 28.12b, this point yields a threshold voltage of $V_T = 0.87$ V, somewhat lower than that determined by the linear extrapolation method ($V_T = 0.9$ V).

The *drain current ratio* method was developed to avoid the dependence of the extracted V_T on the mobility degradation and parasitic series resistance [42]. The drain current is

$$I_D = \frac{W_{eff}\mu_{eff}C_{ox}(V_{GS} - V_T)V_{DS}}{(L - \Delta L) + W_{eff}\mu_{eff}C_{ox}(V_{GS} - V_T)R_{SD}} \tag{28.18}$$

Using

$$\mu_{eff} = \frac{\mu_0}{1 + \theta(V_{GS} - V_T)} \tag{28.19}$$

where μ_0 is the low-field mobility and θ the mobility degradation factor, allows Equation 28.18 to be written as

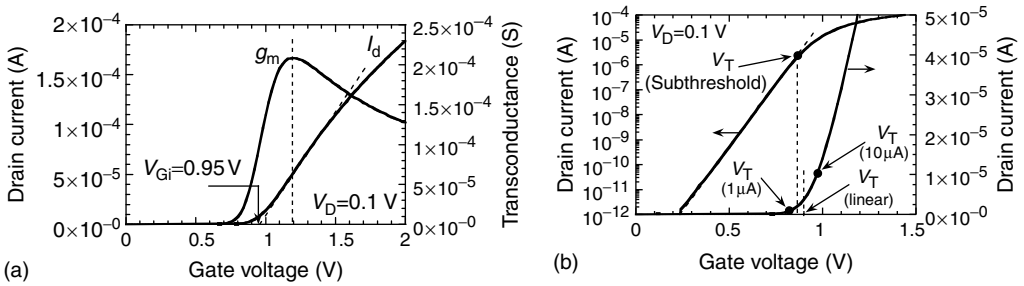


FIGURE 28.12 Threshold voltage determination by the (a) linear extrapolation technique, (b) by the threshold drain current, and the subthreshold techniques, $t_{ox} = 17$ nm, $W/L = 20$ μm/0.8 μm. Data courtesy of M. Stuhl, Medtronic Corp.

$$I_D = \frac{WC_{ox}}{L} \frac{\mu_0}{1 + \theta_{eff}(V_{GS} - V_T)} (V_{GS} - V_T) V_{DS} \tag{28.20}$$

where

$$\theta_{eff} = \theta + (W/L)\mu_0 C_{ox} R_{SD} \tag{28.21}$$

The transconductance is given by

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = \frac{WC_{ox}}{L} \frac{\mu_0}{[1 + \theta_{eff}(V_{GS} - V_T)]^2} V_{DS} \tag{28.22a}$$

The $I_D - g_m^{1/2}$ ratio

$$\frac{I_D}{\sqrt{g_m}} = \sqrt{\frac{WC_{ox}\mu_0}{L} V_{DS}} (V_{GS} - V_T) \tag{28.22b}$$

is a linear function of gate voltage, whose intercept on the gate voltage axis is the threshold voltage. This method is valid, provided the gate voltage is confined to small variations near V_T and the assumptions $V_{DS}/2 << (V_{GS} - V_T)$ and $\partial R_{SD}/\partial V_{GS} \approx 0$ are satisfied. The low-field mobility μ_0 can be determined from the slope of the $I_D - g_m^{1/2}$ vs. $V_{GS} - V_T$ plot and the mobility degradation factor is

$$\theta_{eff} = \frac{I_D - g_m(V_{GS} - V_T)}{g_m(V_{GS} - V_T)^2} \tag{28.23}$$

from which θ can de determined, provided R_{SD} is known.

A good overview of threshold voltage measurement techniques is given in Reference [43]. A comparison of several methods was carried out as a function of channel length [44], showing that the threshold voltage can vary widely depending on how it is measured. In all threshold voltage measurements, it is important to state the sample measurement temperature since V_T depends on temperature. A typical V_T temperature coefficient is -2 mV°C^{-1}, but it can be higher [45].

Strengths and Weaknesses. The strength of the linear extrapolation technique is the common and accepted usage of this method. Its weaknesses are the necessity to differentiate the I_D–V_G curve and fitting of a line, although these steps are automated today. The strength of the threshold drain current method is its simplicity. Its weakness is the choice of the threshold current; different I_T result in different V_T.

28.2.2.2 Effective Channel Length and Source–Drain Resistance

Purpose. The purpose is to determine the MOSFET effective channel length or width and the source/drain series resistance.

Method. The MOSFET current–voltage equation, valid for low drain voltage, is

$$I_D = \frac{W_{eff}\mu_{eff}C_{ox}(V_{GS} - V_T)V_{DS}}{(L - \Delta L) + W_{eff}\mu_{eff}C_{ox}(V_{GS} - V_T)R_{SD}} \tag{28.24}$$

where $W_{eff} = W - \Delta W$, $L_{eff} = L - \Delta L$, V_T is the threshold voltage, W the gate width, L the gate length, C_{ox} the oxide capacitance per unit area, μ_{eff} the effective mobility, and R_{SD} the sum of source and drain resistance. W and L usually refer to the mask dimensions. Equation 28.24 is the basis for most techniques to determine R_{SD}, μ_{eff}, L_{eff}, and W_{eff}. The techniques usually require at least two devices of different channel lengths [46].

In one commonly used method, $R_m = V_D/I_D$ is plotted vs. L as a function of gate voltage for devices with differing L [47]. The lines intersect at one point giving both R_{SD} and ΔL on the R_m and the L axes, respectively. If the R_m vs. L lines fail to intersect at a common point, one can use linear regression to extract both R_{SD} and ΔL [48]. A variation of this method allows ΔL, R_{SD}, μ_0, and θ to be extracted, with μ_0 the low-field mobility and θ the mobility degradation factor [49]. First, R_m is plotted against $1/(V_G - V_T)$, giving slope $m = (L - \Delta L)/W_{eff}\mu_0 C_{ox}$ and intercept $R_{mi} = [R_{SD} + \theta(L - \Delta L)/W_{eff}\mu_0 C_{ox}] = R_{SD} + \theta m$. Then plotting m vs. L, with slope $1/W_{eff}\mu_0 C_{ox}$ and intercept on the L axis of ΔL, allows μ_0 and ΔL to be determined. Lastly, R_{mi} is plotted against m, giving θ from the slope and R_{SD} from the R_{mi} axis intercept.

A technique for any mobility variation with gate voltage and any R_{SD} is the "shift and ratio" method [50]. It uses one large device and several small devices (varying channel lengths, constant channel width). Slope $S = dR_m/dV_G$ is plotted vs. V_G for the large and one small device. One curve is shifted horizontally by a varying amount δ and the ratio $r = S(V_G)/S(V_G - \delta)$ between the two devices is computed as a function of V_G. When S is shifted by a voltage equal to the threshold voltage difference between the two devices, r is nearly constant, which is the key in this measurement. The method has been successfully used for MOSFETs with channel lengths below 0.2 μm. The best range for V_G is from slightly above V_T to about 1 V above V_T. For LDD devices, one should use low gate overdrives to ensure high S allowing dR_{SD}/dV_G to be neglected. Once ΔL is found, R_{SD} can be determined.

Strengths and Weaknesses. The strength of the "R_m vs. L" method is its simplicity. However, the R_m–L lines may not intersect at a single point, pointing to its weakness, especially for LDD devices.

28.2.2.3 Hot Carriers

Purpose. Hot carrier measurements determine the susceptibility of devices to hot carrier (electrons and holes) degradation. Hot carriers are of concern in integrated circuits, because when electrons and/or holes gain energy in an electric field, they can be injected into the oxide to become oxide-trapped charge, they can drift through the oxide, they can create interface-trapped charge, and they can generate photons [51].

Method. The term *hot carriers* is somewhat misleading. The carriers are energetic. The carrier temperature T and energy E are related through the expression $E = kT$. At room temperature, $E \approx 25$ meV for $T = 300$ K. When carriers gain energy by being accelerated in an electric field, their energy E increases. For example, $T = 1.2 \times 10^4$ K for $E = 1$ eV. Hence, the name hot carriers means *energetic carriers*, not that the entire device is hot.

One method to determine hot carrier degradation in n-channel devices is to bias the device at maximum substrate current. The substrate current dependence on gate voltage is shown in Figure 28.13a. The substrate current depends on the channel lateral electric field. At low V_G, with the device in saturation, the lateral electric field increases with increasing gate voltage until $V_G \approx V_D/3 - V_D/2$. I_{sub} increases to a maximum at that gate voltage for n-channel devices. For higher gate voltages, the device enters its linear region, the lateral electric field decreases as does the substrate current. To characterize the device susceptibility to hot carriers, the MOSFET is biased at maximum substrate current $I_{sub,max}$ for a certain time and one device parameter, e.g., saturation drain current, threshold voltage, mobility, transconductance, or interface trap density, is measured [52]. The transconductance is often used. This process is repeated until the measured parameter has changed by some amount (typically 10%–20%), as shown in Figure 28.13b for I_{Dsat}. This time is the lifetime. Next, the substrate current is changed by choosing different gate/drain voltages and the process is repeated and plotted as lifetime vs. I_{sub}, as shown in Figure 28.13c. The data points, measured over a restricted range, are extrapolated to the chip life, typically 10 years, giving the maximum I_{sub}, which should not be exceeded during device operation.

The chief degradation mechanism for n-channel MOSFETs is believed to be interface trap generation and the substrate current is a good monitor for such damage. There are, of course, other measurements that could be used, such as interface trap measurements, for example. I_{sub} is commonly used because it is simple to measure. The main degradation mechanism for p-channel devices is believed to be trapped electrons near the gate/drain interface and it manifests itself at a maximum in gate current. Hence, instead

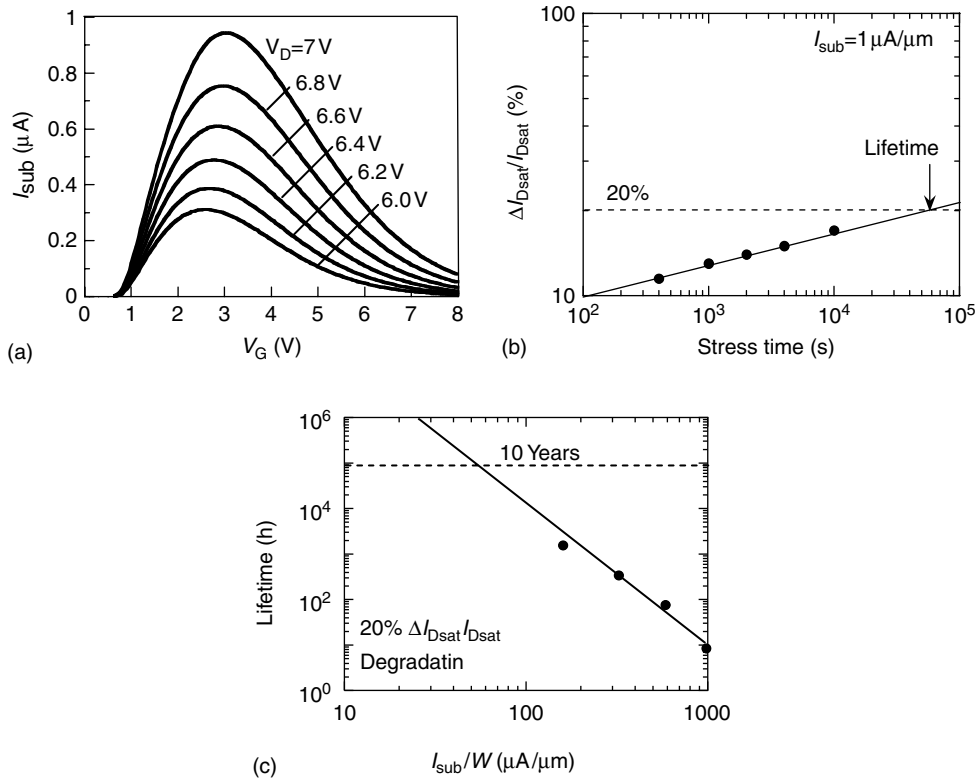


FIGURE 28.13 (a) Substrate current, (b) drain current degradation, (c) lifetime plots for hot carrier degradation. Substrate current plots courtesy of L. Liu, Motorola.

of I_{sub} for *n*-channel devices, in *p*-channel devices I_G is monitored during degradation measurements [53]. Hot carrier damage can be reduced by reducing the electric field at the drain by, for example, forming LDD and using deuterium instead of hydrogen during post-metallization anneal at temperatures around 400°C–450°C, since the Si–D bond is stronger than the Si–H bond [54].

To characterize plasma-induced damage, a common test structure is the *antenna structure* with a large conducting area, consisting of polysilicon or metal layers, attached to a MOSFET or MOS-C gate [55]. Frequently, the antenna resides on a thicker oxide than the MOSFET gate oxide. The ratio between antenna area and gate oxide area has typical values of 500–5000. The antenna test structure is placed into a plasma environment; charge builds up on the antenna and channels gate current through the MOSFET gate oxide where it generates damage that is subsequently detected by measuring the transconductance, drain current, threshold voltage, etc. The highest V_T sensitivity exists for gate oxides 4–5 nm or thicker. Below 4 nm, the gate leakage current is a more suitable measure. Another test structure, the *charge monitor*, is based on an electrically erasable programmable read-only memory structure, consisting of a MOSFET with a floating gate inserted between the substrate and the control gate. The control gate is a large-area collecting electrode [56]. The device is exposed to the plasma, charge builds up, and develops a control gate voltage. Part of that control gate voltage is capacitively coupled to the floating gate. For sufficiently high floating gate voltage, charge is injected from the substrate and is trapped on the floating gate changing the device threshold voltage. The threshold voltage is subsequently measured and converted to charge for a contour map of the plasma charge distribution. The potential sensors are implemented in pairs, where one sensor measures negative and the other positive potentials.

Strengths and Weaknesses. The strength of the I_{sub} measurement is its simplicity. The weakness is that measurements are made under accelerated stress conditions, i.e., when I_{sub} is higher than normal. The resulting data are subsequently extrapolated to normal operating voltages, but the same degradation mechanisms active at high stress may not be active for normal operating conditions.

28.2.3 Oxide Charges, Interface States, and Oxide Integrity

28.2.3.1 Mobile Charge, Interface States

Purpose. The purpose of these measurements is the determination of mobile charge in insulators and interface-trapped charge at the SiO_2/Si interface.

Method. Mobile charge in SiO_2 is due primarily to the ionic impurities Na^+, Li^+, K^+, and perhaps H^+. Sodium is usually the dominant contaminant, but potassium may be introduced during chemical–mechanical polishing and lithium may originate from some pump oils. There are two chief methods: bias-temperature stress (BTS) and triangular voltage sweep (TVS). In the BTS technique, the MOS-C is heated to about 150°C–200°C for 5–10 min with a gate bias to produce an oxide electric field of around 10^6 V/cm. The device is then cooled to room temperature under bias and a *C–V* curve is measured. The procedure is then repeated with the opposite bias polarity. The mobile charge Q_m is determined from the flatband voltage shift, ΔV_{FB}, according to the equation

$$Q_m = -C_{ox}\Delta V_{FB} \tag{28.25}$$

The reproducibility of such measurements becomes questionable as mobile charge densities approach 10^9 cm^{-2}. For example, the flatband voltage shift in a 10 nm thick oxide due to a 10^9 cm^{-2} mobile ion density is 0.5 mV. This is difficult to measure.

In the TVS method, the MOS-C is held at an elevated, constant temperature of 200°C–300°C and both low-frequency (lf) *and* hf *C–V* curves are measured, as shown in Figure 28.14 [57]. Mobile ions are detected as the difference between the two curves [58].

The *quasi-static method* is the most common interface-trapped charge measurement method for MOS-C. One measures the lf and the hf *C–V* curves at room temperature. Interface states are assumed to respond to the lf, but not to the hf curve. In terms of the *measured* C_{lf} and C_{hf}, the interface trap density D_{it} is

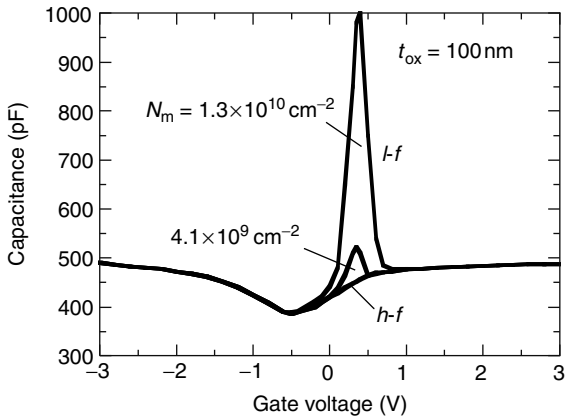


FIGURE 28.14 High-frequency and low-frequency capacitance–gate voltage curves measured at $T = 250$°C. The mobile charge density is determined form the area between the two curves. (After Stauffer, L., Wiley, T., Tiwald, T., Hance, R., Rai-Choudhury, P., and Schroder, D.K., *Solid-State Technol.*, 38, S3–S8, 1995.)

$$D_{it} = \frac{C_{ox}}{q^2}\left(\frac{C_{lf}/C_{ox}}{1-C_{lf}/C_{ox}} - \frac{C_{hf}/C_{ox}}{1-C_{hf}/C_{ox}}\right) \tag{28.26}$$

Equation 28.26 gives D_{it} from the onset of inversion to a surface potential toward the majority carrier band edge where the ac measurement frequency equals the inverse of the interface trap emission time constant. Typically, this corresponds to an energy of about 0.2 eV from the majority carrier band edge. The lower limit of D_{it} determined with the quasi-static technique lies around 10^{10} cm^{-2} eV^{-1}.

Why is $C_{it}=q^2D_{it}$ used here when most texts use $C_{it}=qD_{it}$? $C_{it}=qD_{it}$ is quoted in well-respected texts [59]. But if we substitute units, something is not right. With D_{it} in cm^{-2} eV^{-1} (the usual units) and q in C, the units for C_{it} are

$$\frac{\text{Coul}}{\text{cm}^2\ \text{eV}} = \frac{\text{Coul}}{\text{cm}^2\ \text{Coul}-\text{Volt}} = \frac{F}{\text{cm}^2\ \text{Coul}}$$

using eV = Coul − Volt; Volt = Coul/F. This suggests that the correct definition should be $C_{it}=q^2D_{it}$. We must keep in mind, however, that in the expression $E(\text{eV})=qV$, $q=1$ not $1.6\times q\times 10^{-19}$! Hence, $C_{it}=q^2D_{it}=1\times 1.6\times 10^{-19}D_{it}$. If D_{it} is given in cm^{-2} J^{-1}, then $C_{it}=(1.6\times 10^{-19})^2D_{it}$. This was first pointed out to me by Kwok Ng and can also be found in his book [60].

Interface states in MOSFETs, with gate areas too small for capacitance measurements, are most commonly determined with the *charge pumping technique*. The MOSFET source and drain are tied together and slightly reverse biased. A time-varying gate voltage (square, triangular, trapezoidal, sinusoidal, or tri-level) of sufficient amplitude to drive the surface under the gate into inversion and accumulation is applied. The charge pumping current, measured at the substrate, at the source/drain tied together, or at the source and drain separately, is given by

$$I_{cp} = qAfD_{it}\Delta E \tag{28.27}$$

where A is the gate area and f the gate voltage frequency. The basic charge pumping technique gives an average value of D_{it} over the energy interval ΔE. The energy distribution of the interface traps can be obtained with the tri-level charge pumping technique [61].

Strengths and Weaknesses. The strength of BTS is the simplicity of only hf C–V measurements; its weakness is the inability to distinguish between different mobile species and the time-consuming measurements requiring two heating/cooling cycles. The strength of TVS is the ability to measure more than one species, and measure mobile ions in interlevel dielectrics, with only one heating step. Its weakness is the necessity of measuring both hf and lf C–V curves. The strength of the quasi-static method is the simplicity of measuring the capacitances of MOS-C. Its main weakness is its sensitivity in the low 10^{10} cm^{-2} eV^{-1} and the fact that a rather large-area capacitor must be used. The strength of charge pumping lies in the ability to characterize MOSFETs.

28.2.3.2 Oxide Integrity

Purpose. Oxide integrity measurements are made to determine the breakdown voltage, breakdown electric field, breakdown time, or charge, all of which are indicative of oxide quality.

Method. Oxide integrity is determined by *time-zero* and *time-dependent* measurements. The time-zero method is an I_G-V_G measurement of an MOS device to oxide breakdown. Time-dependent measurements consist of constant or stepped gate current with the gate voltage being monitored, or of constant or stepped gate voltage monitoring the gate current as a function of time. When the oxide is driven into breakdown, one defines a charge-to-breakdown Q_{BD} as

$$Q_{BD} = \int_0^{t_{BD}} J_G \mathrm{d}t \tag{28.28}$$

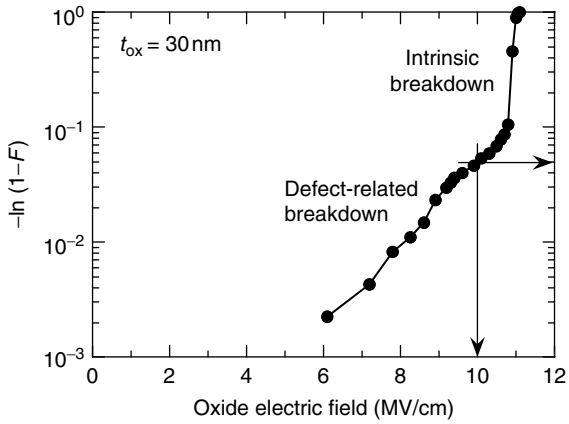


FIGURE 28.15 Cumulative failure vs. electric field. originally published in *Phil. J. Res.*, 40, 1985 (Philips Research).

where t_{BD} is the time to breakdown and J_G the gate current density. Q_{BD} is the charge density flowing through the oxide necessary to break it down. In the stepped current technique, the current is applied for a certain time, e.g., 10 s, it is then increased by a factor of 10 for the same time, etc., until the oxide breaks down [62]. Since, in this method the current starts at a low value, e.g., 10^{-5} A/cm^2, it is a more sensitive technique to bring out B-mode failures, e.g., oxide failures in the intermediate gate oxide electric field range of approximately 3–8 MV/cm.

Oxide breakdown data are usually presented as the number of failures or cumulative failures as a function of oxide electric field or charge-to-breakdown. The statistics of oxide breakdown are described by extreme value distributions [63]. The assumptions underlying the use of extreme value distribution functions are (1) a breakdown may take place at any spot out of a large number of spots, (2) the spot with the lowest dielectric strength gives rise to the breakdown event, (3) and the probability of breakdown at a given spot is independent of the occurrence of breakdown at other spots.

For a device with area A and defect density D, the cumulative failure F is

$$F = 1 - \exp(-AD) \quad \text{or} \quad -\ln(1-F) = AD \tag{28.29}$$

as illustrated in Figure 28.15. Such a plot is known as a Weibull plot [64]. Choosing a particular electric field E_{ox} (in Figure 28.15 it is 10 MV cm^{-1}) gives a value of $-\ln(1-F)$ equal to AD. Hence, this point gives the defect density, provided the area is known, e.g., $-\ln(1-F) = 0.05$, given $D = 5$ cm^{-2} for $A = 0.01$ cm^2.

Strengths and Weaknesses. The strength of constant or stepped current oxide integrity measurements lies in its simplicity and the ease of Q_{BD} extraction. A weakness of typical oxide integrity measurements is the problem all accelerated stress measurements face, namely, is the failure mechanism under accelerated stress the same as that for normal operating conditions. However, since failure under normal conditions takes many years, one is forced into accelerated stress measurements.

28.2.4 Defects and Carrier Lifetimes

28.2.4.1 Deep-Level Transient Spectroscopy

Purpose. Deep-level transient spectroscopy (DLTS) is used to determine energy level, density, and capture cross-sections of deep-level impurities.

Method. For DLTS measurements, the device must be a junction device that is repetitively pulsed into reverse bias and the capacitance, current, or charge is measured as a function of time. The capacitance

is most commonly measured. If the *C–t* curve from a transient capacitance experiment is processed so that a selected decay rate produces a maximum output, then a signal whose decay time changes monotonically with time reaches a peak when the decay rate passes through the rate window of a boxcar averager or the frequency of a lock-in amplifier. When observing a repetitive *C–t* transient through such a rate window while varying the decay time constant by varying the sample temperature, a peak appears in the output plot [65].

The *C–t* transient follows the exponential time dependence

$$C(t) = C_0\left[1 - \left(\frac{N_T}{2N_D}\right)\exp\left(\frac{-t}{\tau_e}\right)\right] \tag{28.30}$$

where C_0 is the steady-state capacitance, N_T the deep-level impurity density, N_D the doping density, and τ_e the emission time constant given by

$$\tau_e = \frac{\exp((E_c - E_T)/kT)}{\gamma_n \sigma_n T^2} \tag{28.31}$$

where E_T is the energy level, γ_n is a constant, and σ_n is the majority carrier capture cross-section.

The *C–t* waveform is sampled at times $t = t_1$ and $t = t_2$ and the capacitance at t_2 is subtracted from the capacitance at t_1, i.e., $\delta C = C(t_1) - C(t_2)$. Such a difference signal is a standard output feature of a double boxcar instrument. The temperature is slowly scanned, while the device is repetitively pulsed between zero and reverse bias. A difference signal is generated when the time constant is on the order of the gate separation $t_2 - t_1$, and the capacitance difference passes through a maximum as a function of temperature. This is the DLTS peak.

A plot of $\tau_{e,max}$ vs. $1/T$ yields E_T and σ_n, while the magnitude of the $\delta C - T$ peaks yields N_T. $\tau_{e,max}$ is obtained from the sampling times t_2 and t_1 as

$$\tau_{e,max} = \frac{t_2 - t_1}{\ln(t_2/t_1)} \tag{28.32}$$

Well-maintained DLTS systems can detect $\delta C_{max}/C_0 \approx 10^{-5}$ to 10^{-4}, allowing impurity densities on the order of 10^{-5} to $10^{-4} N_D$ to be determined. For substrate doping densities of 10^{15} cm^{-3}, one can determine impurity densities to around 10^{11} cm^{-3}. The equipment is commercially available. Energy levels of some of the more important deep-level impurities in silicon are shown in Figure 28.16.

Strengths and Weaknesses. The major strengths of DLTS are the spectroscopic nature of the measurement, allowing species identification, and its high sensitivity. Its major weakness is the need

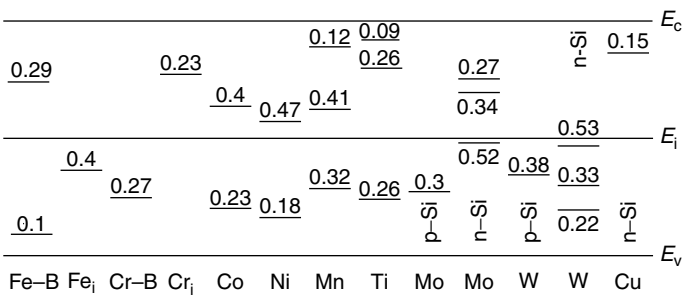


FIGURE 28.16 Energy levels of some impurities in silicon. Energies are indicated for each level. Numbers above E_i are given with respect to E_c, those below E_i with respect to E_v.

for sophisticated measurement equipment including sample cooling. It is also not always possible to assign a unique impurity to a given DLTS spectrum, making impurity identification difficult at times.

28.2.4.2 Recombination Lifetime

Purpose. Recombination lifetimes are most commonly measured to determine impurity density or cleanliness of semiconductors. There is no basic lower limit of impurity density that can be detected through lifetime measurements.

Method. There are many techniques to determine the recombination lifetime. The two most common methods are photoconductance decay (PCD) and surface photovoltage (SPV). During PCD, electron–hole pairs (ehps) are created by optical excitation, and their decay is monitored as a function of time following the cessation of the excitation. This measurement can be contactless by monitoring the reflectance of a microwave signal incident on the sample. The excess carrier density decay for low-level injection is given by $\Delta n(t) = \Delta n(0)\exp(-t/\tau_{r,eff})$ where the effective recombination lifetime, $\tau_{r,eff}$, is a combination bulk, τ_B, and surface lifetime, τ_S, given by

$$\frac{1}{\tau_{r,eff}} = \frac{1}{\tau_B} + \frac{1}{\tau_S}; \qquad \tau_B = \frac{1}{\sigma v_{th} N_T}; \qquad \tau_S = \frac{t}{2\sigma v_{th} N_{it}} \tag{28.33}$$

where σ is the capture cross-section, v_{th} the thermal velocity, N_T the impurity density, and N_{it} the interface trap density. The surface lifetime is a function of surface recombination velocity s_r and sample thickness t. For reasonably low s_r, we can write $\tau_S = t/2s_r$. When τ_B dominates, then a measure of $\tau_{r,eff}$ is a measure of N_T. Hence, a simple recombination lifetime measurement yields information about the level of bulk contamination. For high bulk lifetime material, one can use $\tau_{r,eff}$ measurements to determine the state of the surface, because $\tau_{r,eff} \approx \tau_S$ [66].

The SPV technique yields the minority carrier diffusion length, L, related to $\tau_{r,eff}$ by $L = (D\tau_{r,eff})^{1/2}$, where D is the diffusion constant. An example of using minority carrier diffusion length measurements to determine Fe in Si is shown in Figure 28.17. Iron in p-type Si has the unique property of being in one of two states. When a Fe-contaminated, B-doped Si wafer has been at room temperature for a few hours, the iron forms pairs with boron. Upon heating at around 200°C for a few minutes or illuminating the device, the Fe–B pairs dissociate into interstitial iron, Fe_i, and substitutional boron. The recombination

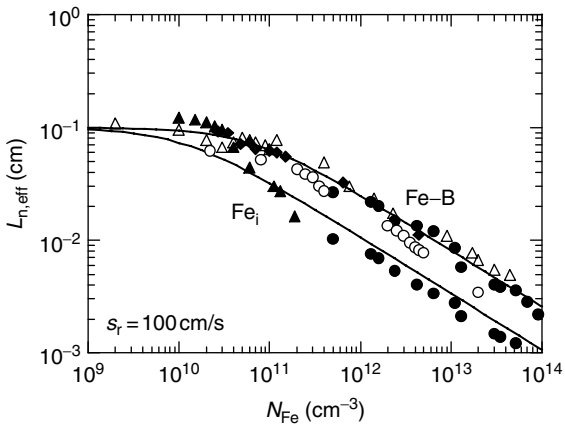


FIGURE 28.17 Effective minority carrier diffusion length vs. iron density in p-type Si. (Data from Schroder, D.K., *Semiconductor Material and Device Characterization*, Wiley, New York, 1998.)

properties of Fe_i differ from those of Fe–B. By measuring the diffusion length of a Fe-contaminated sample before (L_i) and after (L_f) Fe–B pair dissociation, N_{Fe} is [67]

$$N_{Fe} = 1.05 \times 10^{16} \left(\frac{1}{L_f^2} - \frac{1}{L_i^2} \right) \text{(cm}^{-3}\text{)} \tag{28.34}$$

with diffusion lengths in units of microns.

Strengths and Weaknesses. The strengths of both techniques are the contactless nature of the measurements with no sample preparation. The weaknesses are the effect of the sample on the measurement. What is measured is an effective lifetime or diffusion length, which is not always equal to the true value because of sample geometry or surface recombination.

28.2.4.3 Generation Lifetime

Purpose. Generation lifetime is measured to characterize selected regions of a device for contamination. It is especially useful to characterize thin layers, e.g., epitaxial layers, denuded zones, and silicon on insulator (SOI).

Method. The generation lifetime τ_g is related to the recombination lifetime τ_r by [68]

$$\tau_g = \tau_r \exp(|E_T - E_i|/kT) \tag{28.35}$$

Generally, $\tau_g \approx (50\text{–}100)\tau_r$ [41]. The pulsed MOS-C lifetime measuring technique is popular because it is simple and the ubiquitous MOS-C is found on many test structures. The MOS-C is pulsed into deep depletion, and the *C–t* curve is measured. The capacitance relaxes from deep depletion to equilibrium by thermal ehp generation. The effective generation lifetime $\tau_{g,eff}$ is shown in Figure 28.18 as a function of iron concentration, N_{Fe}. $\tau_{g,eff}$ depends on the surface generation velocity s_g. N_{Fe} can be determined from generation lifetime through the relation [69]

$$N_{Fe} = 8 \times 10^8 \left(\frac{1}{\tau_{g,f}} - \frac{1}{\tau_{g,i}} \right) \text{(cm}^{-3}\text{)} \tag{28.36}$$

with τ_g in seconds. $\tau_{g,f}$ and $\tau_{g,i}$ are the final and initial generation lifetimes, i.e., after and before Fe–B pair annihilation.

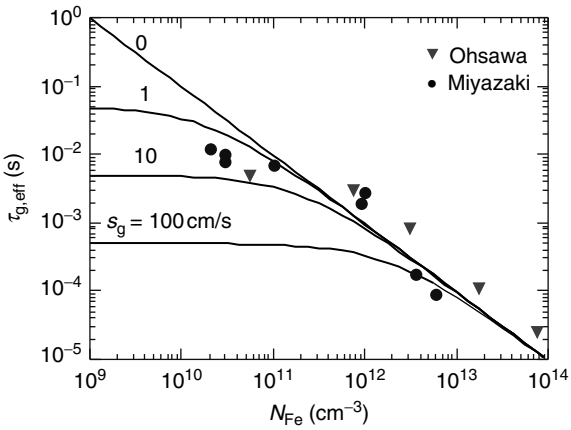


FIGURE 28.18 Effective generation lifetime vs. iron density in *p*-type Si.

The MOS-C pulsed *C–t* technique has found wide acceptance because it is easily implemented with commercial equipment. The measured *C–t* transient times are usually quite long with times of tens of seconds to minutes being common. The relaxation time t_f is related to τ_g by

$$t_f \approx 10\left(\frac{N_A}{n_i}\right)\tau_g \tag{28.37}$$

This equation brings out a very important feature of the pulsed MOS-C technique, which is the magnification factor N_A/n_i built into the measurement. Values of τ_g range over many orders of magnitude, but representative values for silicon devices are 10^{-4} to 10^{-3} s. Equation 28.37 predicts the actual *C–t* transient time to be 10–1000 s. These long times point out the great virtue of this measurement technique. To measure lifetimes in the microsecond range, it is only necessary to measure capacitance relaxation times on the order of seconds. The long measurement times are also a disadvantage because wafer mapping takes a long time, but the times can be reduced by optical excitation [70]. Various pulsed or swept MOS *C–t* techniques have been proposed [71]. A more recent implementation replaces the MOS gate with corona charge, making it a contactless method [72]. It is also possible to determine the generation lifetime from diode leakage current measurements. A particularly good test structure for this purpose is the gate-controlled diode [73].

Strengths and Weaknesses. The strength of generation lifetime measurements is the ease of the measurement and the confinement of the sampled depth to the scr width, easily controlled with the bias voltage. Hence, this technique is well suited for epitaxial layers, denuded zones, and SOI layer characterization. Its major weaknesses are the need for sample preparation, i.e., the sample needs to be oxidized, and the long measurements times.

28.2.5 Charge-Based Measurements

Purpose. Charge-based measurements are rapid "contactless" measurements requiring deposited charge, rather than metal or polysilicon gates. As such, they are rapid measurements providing oxide charge, thickness, and breakdown. They are also used to determine recombination and generation lifetimes.

Method. In charge-based measurements evaporated, sputtered, or deposited, gates are replaced by corona charge and such measurements are used during the development of integrated circuits and for manufacturing control. To be effective, such test structures should provide rapid feedback to the pilot or manufacturing line. Charge, in these measurements, is used in two basic ways: as the "gate" in MOS-type measurements, where the charge replaces the metal or polysilicon gate, and as a surface-modifying method, where the charge controls the surface potential. IBM developed corona charge for semiconductor characterization during 1983–1992 [74]. However, due to lack of commercial instruments, the technique was initially only sparingly used. Later, it was developed into commercial products. During charge-based measurements, charge is deposited on the wafer, as shown in Figure 28.19, and the semiconductor response is measured with a Kelvin probe, first proposed by Kelvin in 1881 [75]. Kronik and Shapira give an excellent explanation of such probes and applications [76].

Deposited charge was first used in the characterization of oxide leakage current and mobile charge drift [77]. Ions are deposited on a surface at atmospheric pressure through an electric field applied to a source of ions. The corona source consists of a wire, a series of wires, a single point, or multiple points located a few millimeter or centimeter above the sample surface. A potential of 5–10 kV of either polarity is applied to the corona source. For a negative source potential, positive ions bombard the source while free electrons are rapidly captured by ambient molecules to form negative ions. Typically, a few seconds are required to charge an insulating surface to a saturation potential.

A surface voltage is generated by the deposited charge or work and is most commonly detected with a non-contacting Kelvin probe. It is a small plate, 2–4 mm in diameter, held typically 0.1–1 mm above the sample and vibrated vertically changing the capacitance between probe and sample at frequencies of typically 500–600 Hz [78].

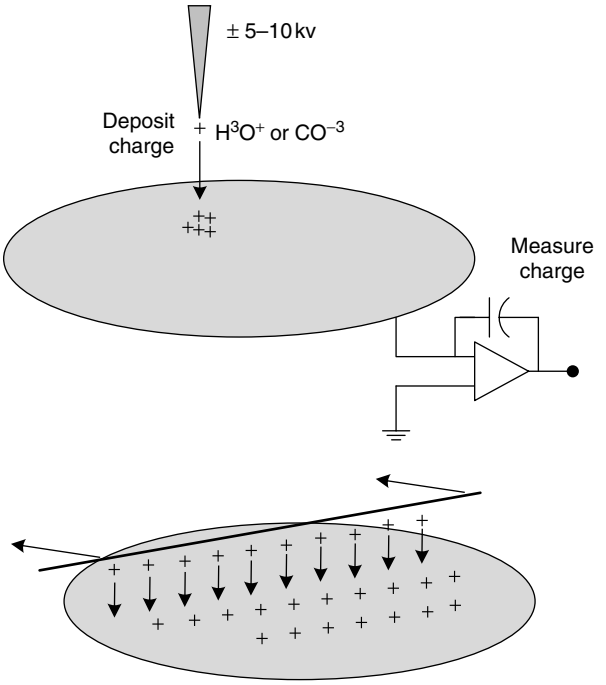


FIGURE 28.19 Schematic illustration of point and wire electrode corona charging methods. The deposited charge is precisely measured with the op-amp charge meter.

The surface voltage dependence on surface charge lends itself to measurements of charge *in* the insulator on a semiconductor wafer or charge *on* the wafer, i.e., oxide charge, interface-trapped charge, plasma damage charge, or other charges.

One determines the surface potential of an oxidized wafer by measuring the surface voltage with and without intense light and deposit corona charge until the surface potential becomes zero. The deposited corona charge is equal in magnitude but opposite in sign to the original oxide charge [79]. The accuracy and precision of this charge-based measurement is identical for thin and thick oxides. Charge-based oxide charge measurements have an advantage over voltage-based measurements. For example, to determine the oxide charge of an MOS device, one can measure the *charge* or the *voltage*. The relationship between the oxide voltage uncertainty ΔV_{ox} and oxide charge uncertainty ΔQ_{ox} is

$$\Delta Q_{ox} = C_{ox}\Delta V_{ox} = \frac{K_{ox}\varepsilon_o \Delta V_{ox}}{t_{ox}} \tag{28.38}$$

Suppose the oxide charge is determined from a voltage measurement with an uncertainty of $\Delta V_{ox} =$ 1 mV. ΔQ_{ox} varies from 2.2×10^{10} to $2.2\times10^{11}\,\text{cm}^{-2}$ for oxide thicknesses from 10 to 1 nm. In voltage-based measurements, there is a large uncertainty in oxide charge. For charge-based measurements, there is a charge uncertainty that is independent of oxide thickness and is on the order of $\Delta Q_{ox}/q = 10^9\,\text{cm}^{-2}$.

To determine the oxide thickness, corona charge density Q is deposited on the oxidized wafer and the surface voltages are measured in the dark and under intense light [80], giving the surface voltage V_s, which is plotted vs. deposited charge density, as shown in Figure 28.20 [81]. In accumulation or inversion, the curves are linear and the oxide thickness is

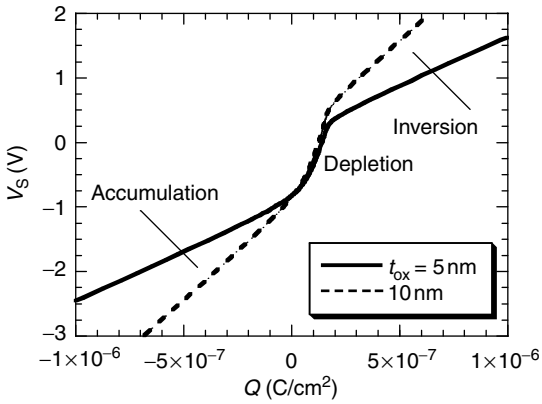


FIGURE 28.20 Surface voltage vs. surface charge density for two oxide thicknesses.

$$C_{ox} = \frac{dQ}{dV_S}; \qquad t_{ox} = \frac{K_{ox}\varepsilon_0}{C_{ox}} = K_{ox}\varepsilon_0 \frac{dV_S}{dQ} \tag{28.39}$$

This method is not subject to the polysilicon gate depletion effects of MOS-C measurements [82]. It is also not affected by probe punchthrough and is relatively insensitive to oxide pinhole leakage currents. Interface traps distort the V_S–Q curve and the interface trap density is determined from that distortion.

For oxide leakage current measurement, corona charge is deposited on an oxidized wafer and the Kelvin probe voltage is measured as a function of time. If the charge leaks through the oxide, the voltage decreases with time. The device is biased into accumulation and the oxide leakage current is related to the voltage through the relationship [83]

$$I_{leak} = C_{ox}\frac{dV_P(t)}{dt} \Rightarrow V_P(t) = \frac{I_{leak}}{C_{ox}}t \tag{28.40}$$

When the device is biased into accumulation, charge builds up on the oxide. However, when the charge density is too high, it leaks through the oxide by Fowler–Nordheim or direct tunneling and the surface voltage becomes clamped. The deposited charge density is related to the oxide electric field ε_{ox} through the relationship

$$Q = K_{ox}\varepsilon_0\varepsilon_{ox} = 3.45 \times 10^{-13}\ \varepsilon_{ox} \tag{28.41}$$

for SiO_2.

Strengths and Weaknesses. The strength of corona-charge-based systems is the contactless nature of the measurements allowing some semiconductor processes to be monitored without having to fabricate test structures as well as the variety of semiconductor parameters that can be determined. A weakness is the specialized nature of the equipment not as routinely found as are current–voltage or capacitance–voltage systems.

28.2.6 Probe Measurements

Purpose. Probe measurements are usually made to obtain microscopic information of a number of material parameters. In the extreme, one can obtain lateral resolution on the order of 0.1 nm and vertical resolution of 0.01 nm.

Method. The SPM refers to techniques in which a sharp tip is scanned across a sample surface at very small distances to obtain two- or three-dimensional images of the surface at nanometer or better lateral and/or vertical resolutions [84]. A chief advantage of probe-based measurements is their high resolution. Other than transmission electron microscopy (TEM), it is the only technique for imaging at atomic resolution. A myriad of SPM instruments has been developed over the past decade, and one can sense current, voltage, resistance, force, temperature, magnetic field, work function, and so on with these instruments at high resolution [85].

Scanning tunneling microscopy (STM) shown schematically in Figure 28.21 [86] consists of a very sharp metallic probe scanned across the sample at distances of about 1 nm, with a bias voltage lower than the work function of the tip or the sample between the tip and the sample. Experimental evidence suggests "mini tips" of <10 nm radii form at the tip of the probes [87]. Piezoelectric elements provide the scanning mechanism. Early implementations used the three-arm tripod arrangement in Figure 9.23, which is subject to low resonance frequencies and was later changed to the tubular implementation containing four symmetric electrodes. Applying equal but opposite voltages to opposing electrodes causes the tube to bend due to contraction and expansion. The inner wall is contacted by a single electrode for actuation voltages for vertical movement [88].

With the probe tip very close to the sample surface, a tunnel current of typically 1 nA flows across the gap. The current is given by [89]

$$J = \frac{C_1 V}{d} \exp\left(-2d\sqrt{\frac{8\pi^2 m\Phi_B}{h^2}}\right) = \frac{C_1 V}{d} \exp(-1.025d\sqrt{\Phi_B}) \tag{28.42}$$

for d in Å and Φ_B in eV, where C_1 is a constant, V is the voltage, d the gap spacing between the tip and sample, and Φ_B an effective work function defined by $\Phi_B = (\Phi_{B1} + \Phi_{B2})/2$ with Φ_{B1} and Φ_{B2} the work functions of the tip and sample, respectively. For a typical work function of $\Phi_B \approx 4$ eV, a gap spacing change from 10 to 11 Å changes the current density by about a factor of 8.

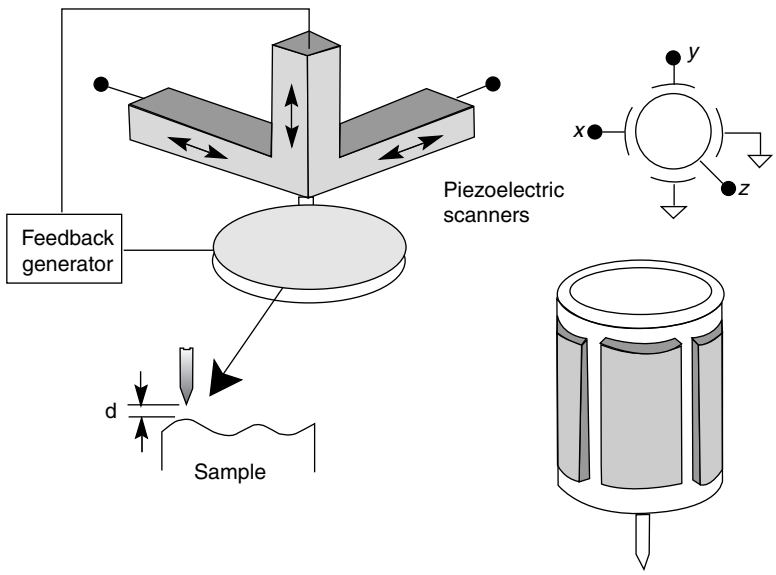


FIGURE 28.21 Schematic illustration of a scanning tunneling microscope.

There are two modes of operation. In the first, the gap spacing is held constant, as the probe is scanned in the x and y dimensions, through a feedback circuit holding the current constant. The voltage on the piezoelectric transducer is then proportional to the vertical displacement giving a contour plot. In the second mode, the probe is scanned across the sample with varying gap and current. The current is now used to determine the wafer flatness. Holding the probe above a given location of the sample and varying the probe voltage gives the tunneling spectroscopy current, allowing the band gap and the density of states to be probed. By using the STM in its spectroscopic mode, the instrument probes the electronic states of a surface located within a few electron volts on either side of the Fermi energy.

The AFM was introduced in 1986 to examine the surface of insulating samples [90]. The AFM operates by measuring the force between a probe and the sample and has evolved into a mature instrument providing new insights into the fields of surface science, electrochemistry, biology, and technology [91]. This force depends on the nature of the sample, the distance between the probe and the sample, the probe geometry, and sample surface contamination.

The AFM principle is illustrated in Figure 28.22. The instrument consists of a cantilever with a sharp tip mounted on its end. The cantilever is usually formed from silicon, silicon oxide, or silicon nitride and is typically 100 μm long, 20 μm wide, and 0.1 μm thick. The vertical sensitivity depends on the cantilever length. For topographic imaging, the tip is brought into continuous or intermittent contact with the sample and scanned across the sample surface. Depending on the design, piezoelectric scanners translate either the sample under the cantilever or the cantilever over the sample. Moving the sample is simpler because the optical detection system need not move. The motion of the cantilever can be sensed by one of several methods [92]. A common technique is to sense the light reflected from the cantilever into a two-segment or four-segment, position-sensitive photodiode, as shown in Figure 28.22 [93]. The cantilever motion causes the reflected light to impinge on different segments of the photodiode. Vertical motion is detected by $z=(A+C)-(B+D)$ and horizontal motion by $x=(A+B)-(C+D)$. Holding the signal constant, equivalent to constant cantilever deflection, by varying the sample height through a feedback arrangement gives the sample height variation. For the beam cantilever, the resonance frequency is given by

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{28.43}$$

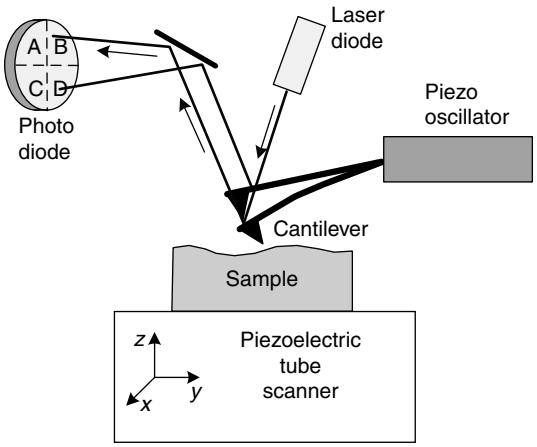


FIGURE 28.22 Schematic illustration of an atomic force microscope.

where k is the spring constant and m the mass of the cantilever. Typical resonance frequencies lie in the 50–500 kHz range.

In the contact mode, the probe tip is dragged across the surface and the resulting image is a topographical map of the sample surface. While this technique has been very successful for many samples, it has some drawbacks. The dragging motion of the probe tip, combined with adhesive forces between the tip and the surface, can damage both the sample and probe and create artifacts in the data. Under ambient air conditions, most surfaces are covered by a layer of condensed water vapor and other contaminants. When the scanning tip touches this layer, capillary action causes a meniscus to form and surface tension pulls the cantilever into the layer. Trapped electrostatic charge on the tip and sample contribute additional adhesive forces. These downward forces increase the overall force on the sample and, when combined with lateral shear forces caused by the scanning motion, can distort measurement data and damage the sample.

In the non-contact mode, the instrument senses van der Waal attractive forces between the surface and the probe tip held above the sample surface. Unfortunately, these forces are substantially weaker than the contact mode forces—so weak in fact that the tip must be given a small oscillation and ac detection methods are used to detect the small forces between the tip and sample. The attractive forces also extend only a short distance from the surface, where the adsorbed gas layer may occupy a large fraction of their useful range. Hence, even when the sample–tip separation is successfully maintained, non-contact mode provides lower resolution than either contact or tapping mode.

Tapping mode imaging overcomes the limitations of the conventional scanning modes by alternately placing the tip in contact with the surface to provide high resolution and then lifting the tip off the surface to avoid dragging the tip across the surface [94]. It is implemented in ambient air by oscillating the cantilever assembly at or near the cantilever's resonant frequency with a piezoelectric crystal. The piezo motion causes the cantilever to oscillate when the tip does not contact the surface. The oscillating tip is then moved toward the surface until it begins to lightly touch or "tap" the surface. During scanning, the vertically oscillating tip alternately contacts the surface and lifts off, generally at a frequency of 50–500 kHz. As the oscillating cantilever contacts the surface intermittently, energy loss caused by the tip contacting the surface reduces the oscillation amplitude that is then used to identify and measure surface features. When the tip passes over a bump on the surface, the oscillation amplitude decreases. Conversely, when the tip passes over a depression, the cantilever has more room to oscillate and the amplitude increases approaching the maximum free air amplitude. The oscillation amplitude of the tip is measured and the feedback loop adjusts the tip–sample separation maintaining a constant amplitude and force on the sample.

FIGURE 28.23 Non-contact AFM image of metal lines showing the grains and grain boundaries. 10 μm × 10 μm scan area. Courtesy of Veeco Corp.

Tapping mode imaging works well for soft, adhesive, or fragile samples, allowing high-resolution topographic imaging of sample surfaces that are easily damaged or otherwise difficult to image by other AFM techniques. It overcomes problems associated with friction, adhesion, electrostatic forces, and other difficulties that can plague conventional AFM scanning methods. An AFM image is shown in Figure 28.23.

In *scanning Kelvin probe microscopy* (SKPM), the probe, typically held 30–50 nm above the sample, is scanned across the surface and the potential is measured. Frequently, this measurement is combined with AFM measurements. During the first AFM scan, the sample topography is measured and during the second scan, in the SKPM mode, the surface potential is determined [95]. The conducting probe and conducting substrate can be treated as a capacitor with the gap spacing being the spacing between the probe and sample surface. The dc and ac voltages are applied to the tip (sometimes the voltage is applied to the sample with the tip held at ground potential). This leads to an oscillating electrostatic force between the tip and sample from which the surface potential can be determined. An advantage of force over current measurements is that the latter is proportional to the probe size while the former is independent of it. The frequency is chosen equal or close to the cantilever resonance frequency, which is typically around several 100 kHz.

An ac voltage of constant amplitude together with a dc voltage is applied. A lock-in technique allows extraction of the first harmonic signal in the form of the first harmonic tip deflection proportional to F_ω. Using a feedback loop, the oscillation amplitude is minimized by adjusting V_{dc}. The detection technique

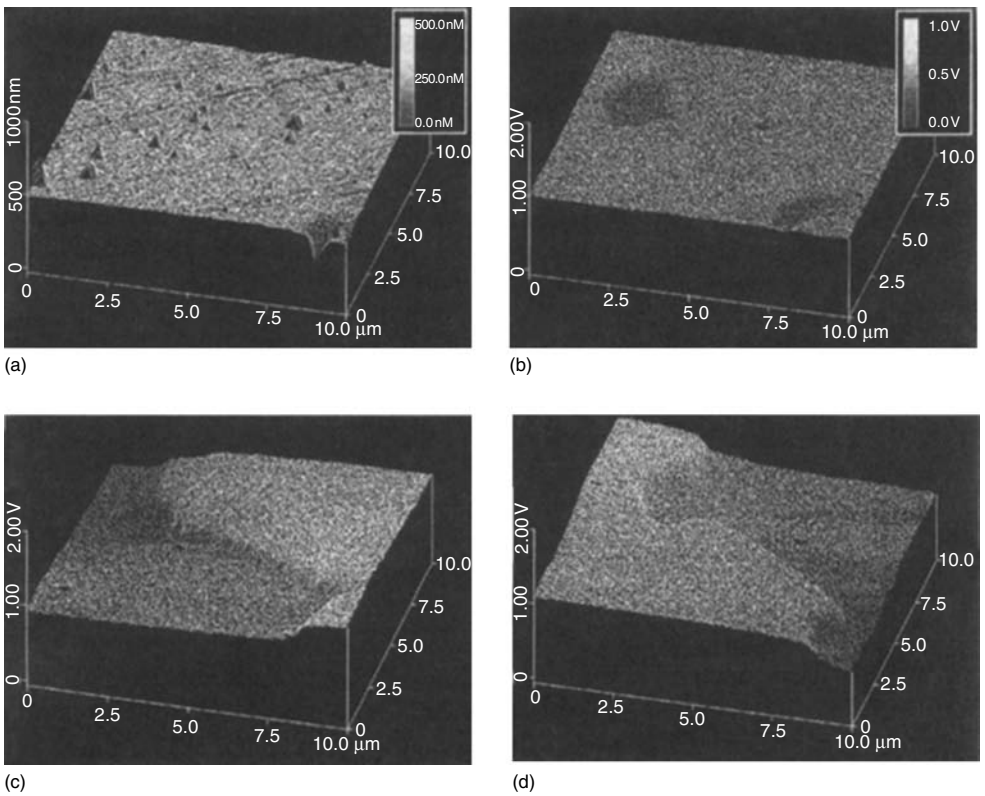


FIGURE 28.24 Polycrystalline ZnO (a) AFM surface topograph, (b) surface potential map with secondary phases and a 100 mV potential difference, (c) positive, and (d) negative voltage applied to the sample showing a 0.3 V potential drop at the grain boundary. The direction of potential drops inverts with bias. (After Bonnell, D.A. and Kalinin, S., *Proceedings of the International Meeting on Polycrystalline Semiconductors*, ed. Bonnaud, O., Mohammed-Brahim, T., Strunk, H.P., and Werner, J.H., 33–47, Scitech Publ. Uettikon am See, Switzerland, 2001.)

is the AFM method with a measure of the feedback voltage V_{dc} being a measure of the surface potential. The null technique renders the measurement independent of dC/dz or to variations in the sensitivity of the system to applied forces. The SKPM has also been combined with optical excitation, similar to the SPV measurements [96]. The AFM and SKPM plots in Figure 28.24 are an effective illustration of surface potentials [97]. The AFM topograph (Figure 28.24a) exhibits no differences associated with multiple phases or grain boundaries in this ZnO sample. In the surface potential map with no external perturbation (Figure 28.24b), a depression of approximately 60 mV is observed due to the difference in work functions of the ZnO surface and pyrochlore phase. The surface potential map with the sample under applied lateral bias shows a potential drop at the grain boundaries in Figure 28.24c and d.

Strengths and Weaknesses. The strength of probe microscopy lies in the variety of possible measurements (topography, electric field, temperature, magnetic field, etc.) and their high resolution to atomic scale. Weaknesses include the measurement time and the fragility of the probes, although recent equipment has become automated and is more rugged than early versions.

28.3 Physical and Chemical Characterization

28.3.1 High Spatial Resolution Imaging

28.3.1.1 Scanning Electron Microscopy

Purpose. The SEM is a versatile instrument for imaging the microstructure of solid surfaces with sub-nanometer spatial resolution [4,98–103]. It is well suited for routine inspections of the intricate details of an integrated circuit in an analytical laboratory or a cleanroom environment. When combined with sectioning and chemical etching techniques, it can be used to delineate stacked metal and oxide films and a variety of crystal defects [104–108]. It is normally configured with an x-ray detector capable of identifying elements of the periodic table in metal films, particles, and generic contamination. Commercial instruments offer 300 mm wafer handling and coordinate driven stages for automated defect review [109].

Method. The basic components are an electron gun, a lens system, scanning coils, an electron collector, and a screen for viewing the image, as shown in Figure 28.25, along with a typical SEM image. Microscopes typically operate in the 1–30 kV range (beam voltage), although there is occasion to use voltages near 1 kV for insulating samples [110]. The condenser and objective lenses focus beam electrons into a small spot, the diameter of which determines the spatial resolution of the microscope, which is near 1.0 nm. Scanning coils deflect the spot in a television-like raster over the surface of the sample. Modern instruments are equipped with field emission electron guns [111] of high brightness, which enhance imaging quality and throughput. The column and sample chamber are evacuated to 10^{-4} Pa (8×10^{-7} Torr) or better.

Image contrast arises from secondary, backscattered, and specimen electrons that are generated as incident electrons penetrate into the sample. The *volume of excitation* depends on the scattering pattern of electrons. Figure 28.26 shows a schematic of the excitation volume and the various secondary signals produced. Penetration depth (R_0) is given approximately by

$$R_0 = \frac{0.0552\ V_0^{1.67}}{\rho} \tag{28.44}$$

where V_0 is the beam voltage and ρ the density of the sample [112]. The low-energy (0–50 eV) secondary electrons (SEs) that are ejected from the near-surface region provide the best lateral resolution. These electrons are also most sensitive to roughness, work function, and charge buildup on the sample. Elastically scattered beam electrons contribute to shadowing effects. All are normally collected by a detector of the Everhart–Thornley type [113] positioned near the sample, or in some designs just above the objective lens. Better imaging resolution can usually be obtained from the upper detector, known as

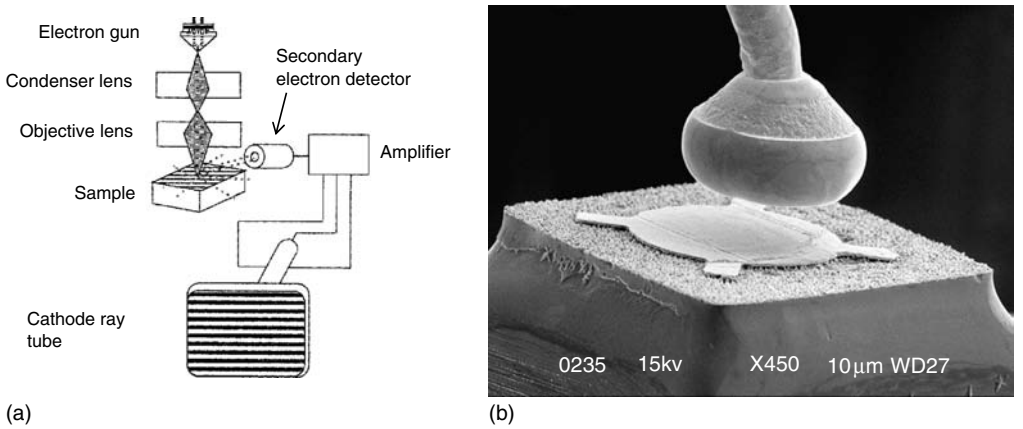


FIGURE 28.25 Basic configuration of a scanning electron microscope (a). Micrograph of bond pad with detached bond wire (b).

the *immersion lens detector* because this detector is effectively shielded from the backscatter electron signal arising from greater depth and, hence, lateral dispersion.

Strengths and Weaknesses. The SEM is easy to use and techniques of sample preparation are straightforward. It is applicable to almost all facets of integrated circuit manufacturing and process development. It also provides a platform for generating and detecting signals other than SE emission, which are generated by the scanning electron beam. This leads to a variety of analytical applications, some of which are listed in Table 28.1.

The most widely used of these is x-ray spectroscopy, which can be acquired from select points on the sample [114–116]. X-ray detectors are of the energy dispersive [117,118] or wavelength dispersive (WDS) [98,116] type, which are optimized for rapid analysis (several minutes) or high spectral resolution ($<$10 eV), respectively. High-energy resolution is useful in quantitative measurements where spectral peak overlaps need to be avoided, as in the determination of phosphorous concentration in a phosphosilicate glass film.

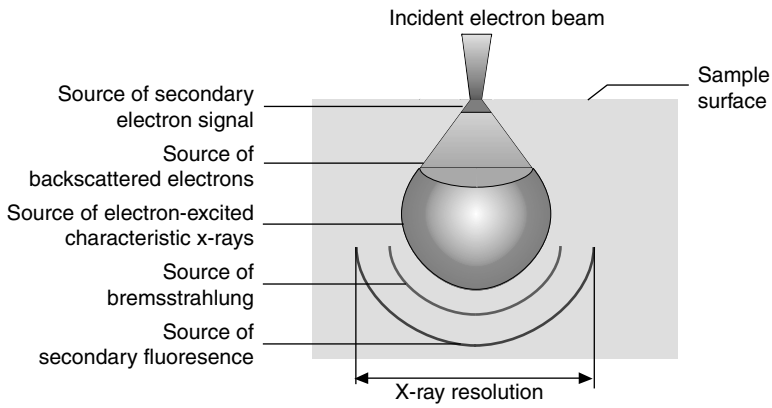


FIGURE 28.26 Diagram showing a cross-section of the excitation volume induced by electron bombardment in a solid. More energetic secondary emissions (backscatter electrons, x-ray emission, fluorescence) can be detected from deeper within the material due to longer mean free paths. At greater depth, the scattering has dispersed the primary beam to a greater degree, compromising the lateral resolutions associated with these signals.

TABLE 28.1 Useful Signals Generated by the Scanning Electron Beam

Technique/Reference	Acronym	Signal Detected	Application Example
X-ray spectroscopy [114–116]	EDS, WDS	0.3–25 kV x-rays	Elemental composition
Auger spectroscopy [119–122]	AES	Auger electrons	Elemental composition
Elemental contrast [98,99]	Z contrast	Backscattered electrons	Heavy metal silicides
Voltage contrast [123–125]	VC	Secondary electrons	Line voltage activity
Electron-beam-induced current [126,127]	EBIC	Electron–hole recombination	pn junction imaging
Cathodoluminescence [128–133]	CL	Visible light	Photonics devices
Kikuchi patterns [134,135]	Kikuchi	Diffracted electrons	Polysilicon and Al grain orientations
Scanning electron acoustic microscopy [136]	SEAM	Electron-induced acoustic waves	Subsurface cracks and voids

X-ray detectors based on Si and Ge crystals are available, the current Ge detectors offering improved spectral resolution. Light element performance can be enhanced by the use of ultrathin windows or windowless detectors. Energy dispersive systems that normally require a liquid nitrogen ambient are available with self-contained Peltier cooling, with some compromise in energy resolution [137]. New microcalorimeter designs promise excellent spectral resolution (approaching 1 eV) and short (several minutes) analysis time [138–144].

It is difficult to use SEM imaging in applications requiring high accuracy, such as CD measurements. This is because extraction of the edge coordinate from an irregular intensity profile is blurred by electron emission enhancement near the edge of a feature. The emission enhancement (edge bright-up), seen at abrupt changes in sample topography, can be reduced by low-voltage operation. The width of lithography features can thus be measured with a high level of precision; the accuracy is usually limited by the calibration methods employed. Automated linewidth measurement systems enhance the reproducibility of such measurements.

The electron beam in an SEM delivers charge to the surface and subsurface region of the sample. In metals and most semiconductors, excess electrons are conducted to ground, but in oxides, nitrides, and packaging materials, charge buildup occurs, giving rise to unpredictable contrasts that complicate image formation and interpretation. Most of this trouble is related to the electric field external to the sample that grows as a consequence of the excess charge [110].

Charging artifacts are controlled by applying a 3 nm metal coating (such as Au, AuPd, Pt, or Cr) over the sample by sputter deposition or thermal evaporation. When continuous and grounded, this effectively eliminates the external field. It is sometimes possible to use a backscattered electron (BSE) detector, which is insensitive to low-energy charging effects, but does not change beam-to-sample interactions. Another approach is to reduce the energy of the primary electron beam, which establishes a charge balance in the absence of a conducting path to ground [110].

Charge balancing can also be achieved in a low-vacuum SEM, in which the sample chamber vacuum can be degraded to a few Torr. In this mode, the column vacuum is protected from the sample chamber vacuum by a differential pumping aperture. Ionization of the gas molecules in the sample chamber automatically charge compensates insulating samples without the need for coating the sample with a conductor or using a low beam voltage. Beam scattering by the chamber gas degrades the instrument resolution and the standard Everhart–Thornley detector cannot be used in this mode.

28.3.1.2 Transmission Electron Microscopy

Purpose. The TEM provides ultrahigh-resolution images of material defects and ultra large-scale integration ULSI device geometry and structure [4,145–153]. With crystalline and polycrystalline materials, it achieves atomic resolution. It is virtually the only tool capable of imaging point defects

related to thermal, mechanical, and implant processing, and is used routinely to measure grain size distributions in polycrystalline materials, as well as the thickness of gate oxides and capacitor composite dielectrics with high accuracy. It is assuming an increasingly important role in semiconductor process development.

Method. Operation of the TEM depends on the sample being very thin. The incident electron beam of the microscope must pass entirely through the sample, which needs to be prepared as a free-standing film only ten to hundreds of nanometers thick. Partially for this reason, the operating beam voltage of the TEM is about an order of magnitude higher than the SEM, and normally falls within the 100–400 kV range.

The typical configuration of a TEM is shown in Figure 28.27. The electron gun is similar to that in an SEM, in which a tungsten filament, LaB_6 cathode, or field emission tip can be used. A condenser lens system shapes the electron beam as it emerges from the gun so that it floods the sample over a broad area. Alternatively, field emission sources are capable of delivering 0.1–1 nA into a probe about 2 nm in diameter. Currents in this range are necessary when performing analytical measurements, such as energy dispersive x-ray spectroscopy (EDS) and electron energy loss spectroscopy (EELS) analyses [103,146,150,151,154]. When one or more of these methods is applied in a materials or device analysis, the practice is referred to as analytical electron microscopy.

The sample is supported on a small copper grid about 3 mm in diameter, which is inserted into a holder positioned at the center of the column. Most holders provide capability for tilt ($\pm 30°$)

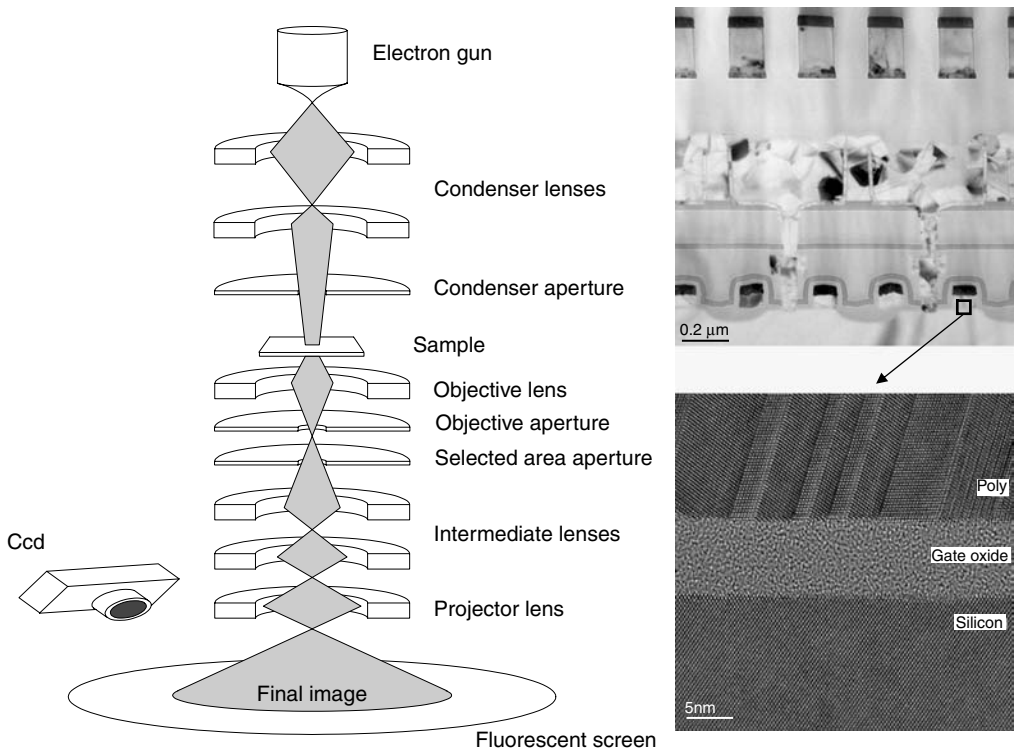


FIGURE 28.27 Schematic diagram of a transmission electron microscopy (TEM) column displaying the path of the electrons through the column. The lower magnification image (upper right) is from a memory array cut along the bit line direction of the array. The high magnification view (lower right) shows the lattice image in the silicon substrate and polysilicon gate from one of the transistors in the array. The lattice image allows for a direct calibration measurement of the gate oxide thickness using the silicon lattice. Data courtesy of Amer Corp.

and translation (± 2 mm) to position a select spot under the beam. Tilt angles up to $\pm 60°$ are routinely available, usually with some loss of resolution because the pole piece gap of the objective lens has to be widened to accommodate the increased tilt angle. An interlock chamber serves to introduce the sample without disturbing the vacuum within the column, which near the stage is in the 10^{-5} Pa (10^{-7} Torr) range.

The objective lens is of short focal length, and produces below it an image of the sample enlarged several hundred-fold or more. This *intermediate image* is further enlarged by a series of intermediate and projector lenses to give a final image, which can be viewed on a fluorescent screen or projected onto a charge-coupled imaging device. In the latter case, the image is directly acquired in digital form and can be processed and archived in a computer database.

The objective aperture (50–200 μm in diameter) is positioned in the back focal plane of the objective lens. It can be moved about during the operation of the microscope, and positioned to block select segments of the electron beam as it spreads out from the underside of the sample. A *selective area aperture* in the first image plane of the system can be used to select a region from which an electron diffraction pattern can be formed; alternatively the incident beam can be made a small diameter and parallel to the optic axis to define a small diffracting region. Elemental and microstructural information can thus be obtained from regions smaller than 20 nm.

Overall, the lenses, apertures, and gun configuration provide working magnifications ranging from a few thousand to several million for viewing the sample. Most semiconductor device analysis can be comfortably carried out at 2000–200,000×. Atomic resolution imaging is more difficult and requires 500,000× or more. In this mode, the technique is called high-resolution transmission electron microscopy.

The basics of electron scattering in the TEM are similar to those for the SEM with two important exceptions. First, the sample is very thin, so the scattering volume is quite small. Second, the higher energy electrons used in the TEM are scattered through angles that are small compared with the SEM, and are almost always deflected in the forward direction.

The angle of scattering (α) determines which electrons are able to pass through the objective aperture, and which are not. Generally, α for electrons that contribute to the image is less than 0.5°, depending on the size of the aperture selected. Electrons deflected by this amount or less contribute to brightness on the screen, while those deflected more are blocked by the aperture, and result in gray or dark regions. If the aperture allows the direct beam to pass, a *bright field* image is formed. If the aperture is positioned to pass one or more diffracted beams and blocks the direct beam, the microscope is operating in the *dark field* mode.

The magnitude of α is determined by the density (ρ) and thickness (t) of the sample. Thickness reduces the intensity of the beam (I_0) according to the expression

$$\frac{I}{I_0} = \mathrm{e}^{-t/\Lambda} \tag{28.45}$$

where Λ is the mean free path between scattering events. In the special case, when $t = \Lambda$, an incident electron on the average will encounter only one scattering event. Λ has been determined experimentally, with Si being near 0.12 μm, and metals like Cr, Ge, Pd, and Pt about an order of magnitude smaller [155]. The thickness of most cross-sections prepared for TEM is of the order of Λ; for atomic resolution imaging, it must be in the tens of nanometer range to give high-quality images.

Many TEM instruments are fitted with scan coils so that the illumination can be rastered over the sample. The pre-specimen lenses in these instruments can be configured to provide a small focused spot, < 0.5 nm for field emission sources. The transmitted beam falls on a detector and is used to modulate the brightness of a display CRT scanning in synchrony with the beam. The resulting image is a scanning transmission image and the technique is referred to as scanning TEM. Image contrast is formed via a post-specimen aperture as in conventional TEM.

Strengths and Weaknesses. The distinctive advantage of the TEM is exceptionally high spatial resolution, even atomic resolution, which can be achieved in nearly routine applications. An example

TABLE 28.2 Application of TEM Sample Preparation Procedures

Technique	Semiconductors	Metals	Organics
Dimple polish	F	F	R
Chemical thinning	F	F	R
Ion milling	F	O	R
Focused ion beam	F	F	O
Tripod select area	F	R	R
Jet polishing	O	F	R
Cleaving	O	O	R
Replica casting	R	F	F
Ultra-microtomy	R	R	F

F, frequently used; O, occasionally used; R, rarely used.

is shown in Figure 28.27. Capabilities can be expanded beyond direct imaging by collecting signals other than electron scattering, which are simultaneously generated within the sample. These include EDS, EELS, electron diffraction, and electron holography.

The challenge with TEM is sample preparation. Techniques typically involve a combination of mechanical polishing, chemical etching, and ion beam milling. Most reflect the experience of the individual operator who learns through trial and error which procedure works best for the material or device at hand. Some generic approaches developed over the years include those listed in Table 28.2. References for all of these are available in a series of volumes on TEM sample preparation published by the Materials Research Society [156–158].

28.3.1.3 Focused Ion Beam Sample Preparation

Purpose. The focused ion beam (FIB) tool uses a narrow pencil of Ga^+ ions to selectively carve from the bulk sample a cross-section suitable for optical microscopy, SEM or TEM imaging, and microanalysis. It is capable of cutting through metal and polysilicon interconnects as well as oxide and nitride layers neatly and precisely, with little or no damage to adjacent structures. Via holes can be created to expose underlying metal landing pads for mechanical probe contact. It is possible to routinely expose a micro-defect buried deep within select transistors in a multi-megabit array or prepare a smooth surface for polycrystalline grain size imaging enhanced by ion channeling. Modern instruments accommodate production size wafers and are cleanroom compatible.

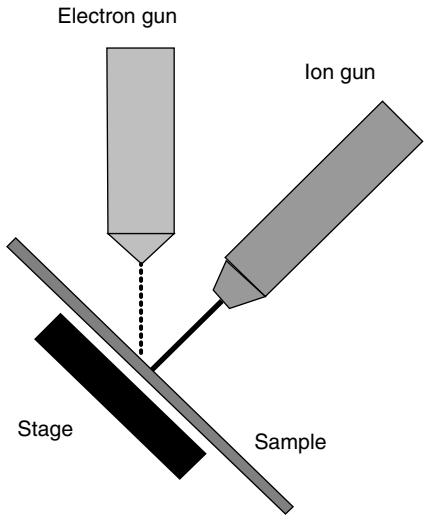
Method. The FIB instruments use a finely focused probe of Ga^+ ions to etch any small region identified at the surface of an integrated circuit [4,159–162]. Typically, a high-current broad beam (0.5 μm diameter) is applied for an initial rough cut, followed by a tighter focus (100–10 nm) low-current probe for final polishing. Ion-induced electron images resemble SEM images, with the exception that channeling effects are pronounced in crystalline and polycrystalline materials [163].

TABLE 28.3 Comparison between Scanning Electron Microscope (SEM) and Focused Ion Beam (FIB) Microprobe Columns

	SEM	FIB
Beam	Electrons	Ga^+ ions
Source	Tungsten, LaB_6 or field emission tip	Field emission liquid Ga^+ tip
Beam energy	100 V–30 kV	5–50 kV
Beam current	>30 nAmp	1–10 nAmp
Minimum spot	<1.5 nm	<10 nm
Lens design	Magnetic	Electrostatic
Surface effects	Secondary and backscattered electrons	Sample sputtering and secondary electron generation
Etch rate	None	$2\ \mu m^3/s \cong 10$ nA

Ga^+ ions are extracted from a liquid droplet in the ion gun by an intense electric field, similar to the way electrons are drawn from a field emission tip in an SEM. The tip of the liquid is only 100 nm across, so it is possible to demagnify this into a probe less than 10 nm in diameter at the sample's surface. The basic components in the FIB include lenses, defining apertures, and scanning coils to raster the probe across the sample. A comparison between the SEM and the FIB is given in Table 28.3.

The larger spot of the Ga^+ probe results in lower resolution than that possible in the SEM, but for most applications it is adequate to locate the region of interest for milling. Nevertheless, most modern FIB vendors offer an independent SEM column as an accessory. These *dual beam* instruments are more costly, but provide the best possible imaging resolution, even in situ while milling takes place. Figure 28.28 shows a schematic of a dual beam FIB along with an SEM image of a cross-section produced by the instrument. Cross-section analysis of such high aspect ratio vias would be difficult and



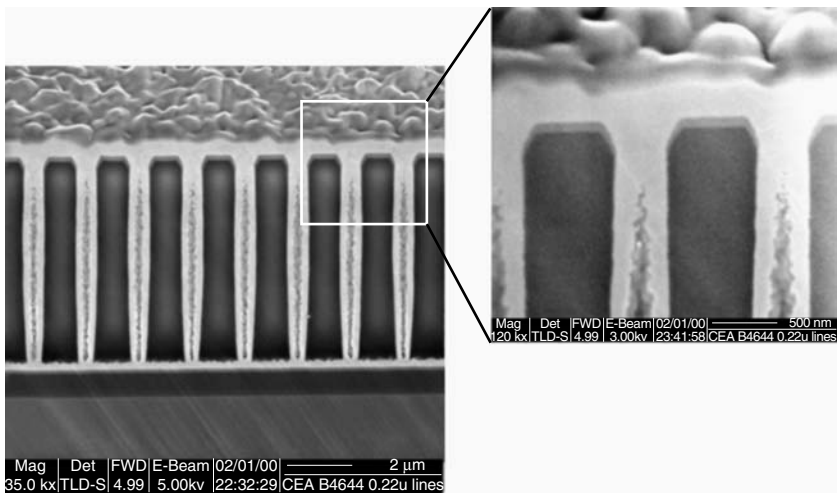

Mag Det FWD E-Beam 02/01/00 2 µm
35.0 kx TLD-S 4.99 5.00kv 22:32:29 CEA B4644 0.22u lines
Mag Det FWD E-Beam 02/01/00 500 nm
120 kx TLD-S 4.99 3.00kv 23:41:58 CEA B4644 0.22u lines

FIGURE 28.28 Diagram of a dual beam focused ion beam (FIB) instrument. The images show low and high magnification views of a FIB cross-section through high aspect ratio vias. Voids in the metal fill are evident.

TABLE 28.4 Chemistries for Assisted FIB Etching

Etch Gas	Al	W	Si	SiO_2
Cl_2	X		X	
Br_2	X		X	
ICl	X	X	X	
ClF	X	X	X	
XeF_2		X	X	X
WF_6				X

time consuming using standard SEM sample preparation methods; however, the FIB can accomplish tasks of this scope within a couple of hours or less.

It is possible to enhance the etch rate of select materials by bleeding a small quantity of gas through a small jet near the sample while milling is in progress. Some of the gases used for microcircuit applications are shown in Table 28.4.

Strengths and Weaknesses. Milling with the focused Ga^+ ion beam provides unique capability for preparing a cross-section at any predetermined spot on an integrated circuit with <10 nm polishing precision. Applications to single bit sectioning in a dynamic random access memory array, for example, is extremely difficult by any other approach.

It is also possible to deposit metals such as W, Pt, or Au within the same FIB machine with the aid of the Ga^+ beam [164,165]. This can be done with submicron resolution to connect two lines together, or to provide bond pads for offline mechanical probing of small or buried structures. The instrument serves a dual role in this regard. It can be used to prepare select cross-sections by precision milling, as well as rewire intricate circuits for electrical probing or SEM voltage contrast imaging.

Most of the Ga^+ ions scatter into the environment, but a significant number are implanted into the sample, which amorphizes the outer 30 nm of crystalline samples. Also, Ga^+ injection is known to alter the electrical properties of single transistors [166]. For production wafers, Ga^+ contamination is of concern, particularly if they are to be inserted back into the process line.

The basic chemistry of the deposition process is complex and not very well understood. Deposited films typically contain impurities like carbon from the vacuum, which can result in a hard, brittle alloy with high electrical resistivity [167]. New gas precursors for depositing oxides and thin dielectrics are being evaluated [168].

28.3.1.4 Scanning Probe Microscopy

Purpose. The SPM is able to provide images of single atoms at the surface of a sample and can also be used as an ultra-sensitive profilometer for measuring roughness. The latter is the most frequently used for semiconductor process and development applications. While the conventional stylus profilometer may sense vertical relief differences as small as 10 nm, the scanning probe easily extends into the 0.01 nm regime when required. Hence, it is a unique tool for assessing roughness related to silicon wafer polishing, implant damage, plasma etching, chemical cleanups, and a variety of other process operations.

Method. The SPMs function by positioning a very sharp needle within fractions of a nanometer of the surface of a sample using piezoelectric micro-manipulators. The two most common forms of SPM (STM and AFM) were described in detail in Section 28.2.6, earlier in this chapter. Similarly, SCM and SSRM were described in Section 28.2.1.4.

There are many other variations to the SPM theme, all of which ultimately depend on the design and function of the tip itself. For example, a tiny thermocouple fabricated just at the apex of the probe is the basis for the scanning thermal microscope, while a tapered quartz fiber channels visible light through a 20 nm aperture in the near-field scanning optical microscope. This list can be readily expanded to include a dozen or more other techniques, all of which are covered in a growing number of articles, books, conferences, and specialized symposia [8,169–175]. Some of them are listed in Table 28.5.

TABLE 28.5 Variations of Scanning Probe Techniques

Acronym	Technique	Reference
AFM	Atomic force microscopy	[173]
BEEM	Ballistic electron emission microscopy	[176]
CFM	Chemical force microscopy	[177]
IFM	Interfacial force microscope	[178]
MFM	Magnetic force microscopy	[179]
MRFM	Magnetic resonance force microscopy	[180,181]
MSMS	Micromagnetic scanning microprobe system	[182]
Nano-NMR	Nanometer nuclear magnetic resonance	[183]
Nano-Field	Nanometer electric field gradient	[184]
Nano-SRP	Nanometer spreading resistance profiling	[185–187]
NSOM	Near-field scanning optical microscopy	[188–190]
SCM	Scanning capacitance microscopy	[191–193]
SCPM	Scanning chemical potential microscopy	[194]
SEcM	Scanning electrochemical microscopy	[195]
SICM	Scanning ion-conductance microscopy	[194]
SKPM	Scanning Kelvin probe microscopy	[196]
SThM	Scanning thermal microscopy	[172]
STOS	Scanning tunneling optical spectroscopy	[197]
STM	Scanning tunneling microscopy	[173]

Strengths and Weaknesses. The AFM is sensitive enough to trace out the contours of single-surface atoms on the sample in a matter of minutes, but as with the STM, it is more frequently applied in industrial laboratories for surface roughness mapping (Figure 28.29). In this mode, the X and Y scan ranges are typically limited to 100–200 μm or less, while the vertical (Z) retains sensitivity to sub-nanometer changes over the scanned area. Instruments suitable for the cleanroom environment accommodate 200 mm wafers, and feature automated cassette loading and inspection.

The application of STM to practical problems is limited to conducting samples, which are capable of carrying electrons away from the tip and surface. Even with sub-nanoampere tunneling currents, charge buildup in thin surface oxides or contamination films is enough to mask or destroy the signal. Materials like Si, Al, Ti, and W readily oxidize and are frequently difficult to image reproducibly. Models that explore the parameters governing the tip and sample confirm that the situation is quite complex [198–200]. This is particularly true when operating in an atomic resolution mode.

Practical knowledge about tips is based on experience with their fabrication and use. A tungsten tip prepared by electrochemical etching with a perfectly smooth end of small radius does not always provide

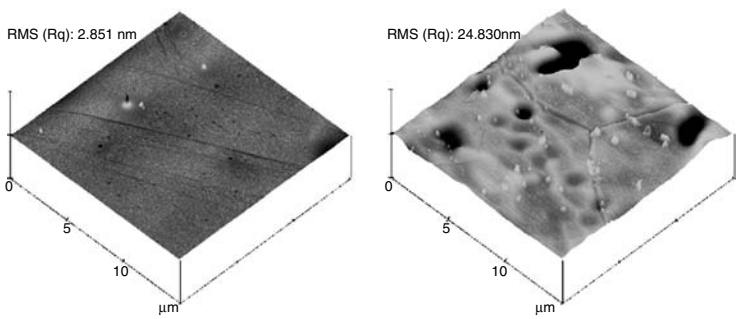


FIGURE 28.29 AFM scans comparing the surfaces of stainless steel tubing before (right) and after electropolishing (left).

atomic resolution at first. Tips that do, often give unpredictable and non-reproducible tunneling current curves. Removal of oxides by annealing or exposure to high field strength before and during operation is usually beneficial. Tips fabricated from Pt to Ir wire may work well for tunneling spectroscopy, even when extremely dull at the end. Serious CD measurements are limited by the shape of the tip itself. Attempts to reconstruct the tip silhouette during analysis are being made, with the intent of removing the influence of shape by mathematical deconvolution [201,202].

28.3.2 Dopants and Impurities

28.3.2.1 Secondary Ion Mass Spectrometry

Purpose. Secondary ion mass spectrometry (SIMS) offers a unique combination of small analytical spot and high detection sensitivity and measurement precision, enabling the technique to monitor dopants and impurities within a patterned contact or junction. At high material removal (sputter) rates, SIMS is especially well suited for characterization of dopant distributions prepared by diffusion or ion implantation with typical detection limits of or below 10^{15} atoms per cubic centimeter. The technique is capable of reaching parts per million atomic (ppma) sensitivity within a contact or junction only 1–10 μm across. SIMS has been the workhorse of the industry for verifying implanter performance, implanter dose matching, and anneal processes, because it can readily determine the dose and shape of the dopant profile, as well as reveal metal impurities. Given appropriate and carefully controlled analysis protocols, dose measurement precisions well below 1% can be routinely achieved. Using low sputter rates and appropriate spectrometer configuration, SIMS can also provide high-sensitivity surface metal and organic contamination analysis in 50×50 μm^2 areas at sub-nanometer analysis depths.

Numerous books, conferences, and workshops specializing in SIMS address the wide range of applications and technique improvements, as well as its limitations [4,203–207,218–220].

Method. All SIMS tools employ a focused primary ion beam of appropriate kinetic energy to strike the analysis site, causing the emission of charged and (mostly) neutral secondary atoms and molecules from the bombarded surface. The charged particles (secondary ions) of selected polarity are injected into the analyzer for mass separation. The emission of charged and neutral particles due to the primary ion impact obviously involves an erosion of the analysis site. For constant primary ion energy, ion species and impact angle on the sample surface, the erosion rate is proportional to the rate of primary ions striking the surface during analysis (i.e., the primary ion current). The use of high primary ion current densities, while monitoring appropriate secondary ion masses, allows for fast surface erosion rates and is ideally suited for profiling beneath the surface of the sample. At the other extreme, highly surface-sensitive analysis is usually performed using very low sputter rates of a few nanometer per hour, allowing a survey of inorganic, metal, or organic contaminants from the top few monolayers of a single analysis site.

Although SIMS can in principle be performed with almost any energetic ionized element of the periodic table, only a few are employed in routine measurements. For depth profiling applications, the most commonly used primary ion species are O_2^+ and Cs^+ at impact energies from 250 eV to 15 keV, depending on the analysis task. The sensitivity of a SIMS measurement is strongly affected by the ion bombardment conditions and the composition of the matrix. In general, non-reactive ion beams such as Ar^+, Xe^+, and Ga^+ produce secondary ion yields from matrices, which are orders of magnitude lower than those measured with reactive beams such as O_2^+ or Cs^+ [207–209]. For this reason, SIMS instruments generally have two separate primary ion columns to produce sputter beams: O_2^+ and Cs^+ for the analysis of elements that are likely to form positive and negative secondary ions, respectively (electropositive and electronegative elements). Positive secondary ion analysis using O_2^+, Ga^+, or other inert primary ion beams in conjunction with O_2 gas flooding is also often employed to enhance secondary ion formation and to reduce or eliminate surface roughening in ultra-shallow depth profile applications.

As a general rule, the analysis of shallow structures requires the use of low-energy primary ion beams to minimize ion-beam-induced mixing of the sample and to improve depth resolution. It was recognized early on that high primary ion bombardment energies skew the shape of implant profiles more than lower bombardment energies [210–212]. In addition, it was demonstrated [213] that the primary ion

bombardment results in essentially isotropic mixing of surface and subsurface layers. The expected broadening of structures toward and away from the surface was calculated to be on the order of the primary ion range [214], in reasonable agreement with measured decay lengths [215].

For example, a 500 eV boron implant into silicon has a projected range of around 3.6 nm and a range straggle of 1.9 nm [216]. Ultra-shallow depth profile analyses of such implants are therefore generally performed using $\leq$500 eV O_2^+ primary ion beams.

In general, secondary ions of at least the dopant and a signal corresponding to the substrate (matrix signal) are measured as a function of time during the depth profile. The ratio of the measured dopant signal strength to the strength of the matrix signal is then converted to a dopant concentration by applying a normalization factor called a "relative sensitivity factor" (RSF). Similarly, the profile shape of the dopant is derived by additionally applying a calibrated/measured sputter rate to the measured data curves.

Sufficient mass separation of the secondary ion signals is required so that the ion species monitored during a depth profile are in fact due to the dopant. The ability to separate two atoms or molecules of the same charge at mass m, which differ in mass by the amount Δm is described by the mass resolution as

$$\text{mass resolution} = \frac{m}{\Delta m} \tag{28.46}$$

A higher mass resolution value of a spectrometer implies better ability to unambiguously measure and identify the element or molecule of interest. For example, the analysis of ^{11}B in silicon requires only a mass resolution of 11 because the closest interfering signal is due to ^{12}C. Higher mass resolution is required for separating interferences between species like $^{31}P^+$ and $(^{30}SiH)^+$ or $(^{28}Si_2)^+$ and $^{56}Fe^+$, which occur in the SIMS analysis of silicon.

The most common mass spectrometer types employed in depth profile analysis are *magnetic sector* instruments of the Nier–Johnson geometry, as well as *quadrupole* mass analyzers. Both instrument types employ continuous primary ion beams and can detect only one secondary ion species at a given instant during the depth profile. In practice, at least two secondary ion species are monitored during a depth profile by periodically switching the analyzer to detect the secondary ion species of interest. Because while one mass is monitored, all other signals of interest are not detected, the ultimate detection sensitivity will decrease with the number of ion species monitored during the profile. Magnetic-sector-based SIMS spectrometers can provide mass resolution in excess of 40,000, whereas quadrupole SIMS spectrometers operate at $m/\Delta m$ of 350. For applications requiring high mass resolution (e.g., P in Si) magnetic sector SIMS is obviously the instrument of choice. In situations where low mass resolution is acceptable (e.g., ultra-shallow B in Si), Quadrupole SIMS instruments can often match or outperform their magnetic sector counterparts, in part because of the relative ease in utilizing sub-kilo-electronvolt primary ion beams at various impact angles for ultra-shallow depth profile analysis and also for their simplicity of operation. Only the most recently developed magnetic sector instruments are able to perform depth profile analysis with sub-kilo-electronvolt primary ion beams. The range of depth profile analyses may be anywhere between a few nanometers to several microns.

Although originally focused toward surface analysis of the topmost monolayer(s), Time-of-flight (ToF) SIMS spectrometers have lately been employed in depth profiling applications. Time-of-flight-SIMS employs a pulsed primary ion beam of nanosecond duration to strike the surface, generate secondary ions, and transport them through an electrostatic analyzer to the detector. Because all secondary ions basically travel along the same path through the analyzer and mass separation is solely due to flight time differences from the sample to the detector, a ToF analyzer is capable of detecting *any* secondary species of given polarity over a mass range 1 kDa to some 10 kDa and $m/\Delta m$ up to 15,000. As a consequence, ToF-SIMS has unsurpassed overall (parallel) detection sensitivity per surface layer. The material removal rate using a pulsed primary ion beam alone is so slow (some nanometer per hour) that the depth profile analysis requires the additional alternating use of a sputter ion beam (10 s of μs in duration) to advance in depth. None of the material removed by the sputter beam can be mass analyzed. A ToF-SIMS depth profile is in essence a sequence of complete surface analyses at increasing depth from

the surface, i.e., the depth profiles of all detectable ions species are collected in a single analysis. Depending on the mode of operation, the depth range of analyses covered with ToF-SIMS is between the surface to (reasonably) some 100 nm.

For details on mass spectrometer types, ion beam scanning and crater edge rejection methods, etc., the reader is referred elsewhere [207,217].

Secondary ion mass spectrometry analysis performed at "high" sample erosion rate is often referred to as *dynamic* SIMS, whereas the other extreme, where erosion rates of a few nanometer per hour are employed is often called *static* SIMS and was originally applied to organic surface analysis. Today's SIMS depth profiling applications cover bulk dopant analysis, high, and ultralow energy ion implant characterization, gate dielectric characterization, and surface metal contamination analysis. The terminology has therefore become somewhat arbitrary—especially when applied to distinguish dedicated depth profiling instruments from ToF-SIMS analyzers. In practice, the purpose of the SIMS analysis will first dictate the primary ion bombardment conditions (erosion rate, ion species, energy, impact angle, and beam size) and then which type of SIMS instrument is best suited to provide the primary ion beam conditions *and* the required mass resolution, detection sensitivity, data collection rates, precision, etc. For example, the quantification of specific bulk dopants or high- and medium-energy ion implants requires large analysis depth and thus the high erosion rates provided by high-energy ion beams that are well accommodated by dedicated *magnetic sector* (or *quadrupole*)-based depth profiling spectrometers to

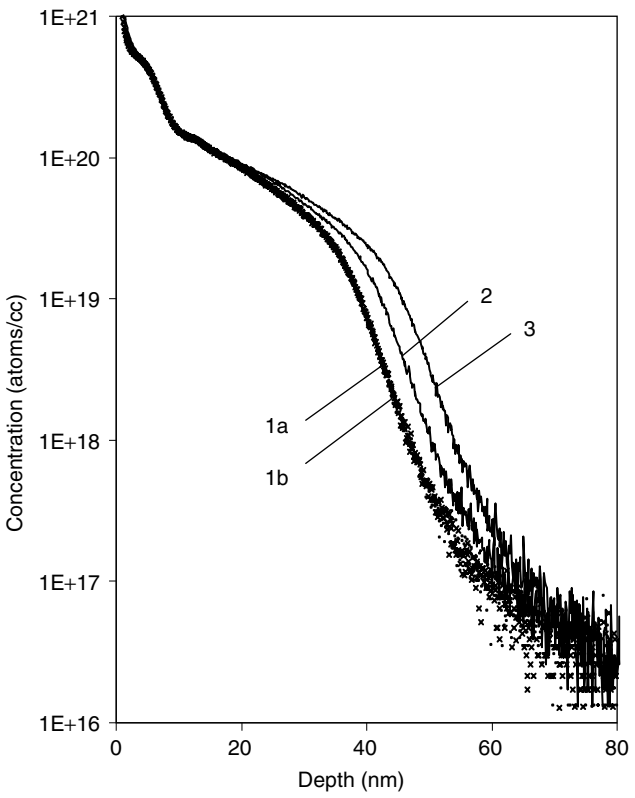


FIGURE 28.30 Depth profiles of B implants into Si, processed by three different spike anneals (1, 2, and 3) to investigate transient enhanced diffusion. The depth profiles of process 1 were taken before (1a) and after (1b) the depth profiles of wafers 2 and 3, demonstrating $<1\%$ reproducibility/stability of the measurements of a *quadrupole* Secondary Ion Mass Spectroscopy (SIMS) instrument.

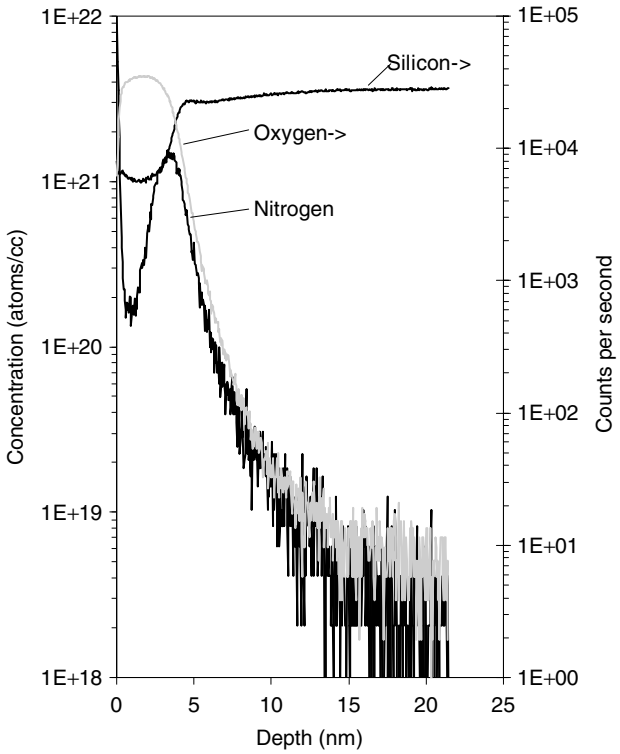


FIGURE 28.31 Depth profile of a 3.7 nm thick oxynitride film, obtained on a *quadrupole* SIMS instrument, using 500 eV Cs^+ primary ion bombardment. The nitrogen dose of the film is 3.7×10^{14} atm/cm^2. Nitrogen is primarily located at the Si substrate interface, as expected from the film processing conditions.

achieve ppma detection limits. On the other hand, the characterization of ultra-shallow implants requires the use of low-energy (500 eV–2 keV) primary ion beams, readily available in modern *quadrupole* SIMS instruments and only recently realized in the newest magnetic-sector-type spectrometers. The primary applications are low-energy implant dose matching, characterization of anneal processes (Figure 28.30), as well as characterization of ultrathin films such as gate oxides (Figure 28.31). Ultra-shallow depth profiling is also possible with ToF-SIMS instruments (but will *not* improve the depth resolution) if equipped with suitable dedicated sputter guns. Finally, the characterization of surface metal contamination caused by the ion implanter or qualification of cleaning processes requires high detection sensitivity (ppm or better), as well as high mass resolution, which can be readily achieved with magnetic sector or ToF-SIMS instruments.

The different analytical and instrumentation approaches offer flexibility for achieving the best performance possible in the analysis of submicron circuit junctions, oxide composite films, and surface organic contamination. A comparison of SIMS techniques is given in Table 28.6.

Strengths and Weaknesses. SIMS is one of the most destructive surface analytical techniques, but it is also one of the most sensitive and reproducible techniques available. This applies to most elements in the periodic table, including H, C, N, and O, although these are also common contaminants found in the vacuum environment. The SIMS can reproducibly detect sub-% dose and in-depth distribution variations in low- and high-energy ion implants, as well as ultrathin films.

Analysis on small areas is available with all instrument types discussed, as long as sufficiently well-focused and intense primary ion beams can be produced (microprobe). Most commonly employed magnetic-sector-type spectrometers are designed to provide microns spatial resolution with either tightly

TABLE 28.6 Application Areas of Different Secondary Ion Mass Spectroscopy (SIMS) Techniques Based on Different Instrument Design Configurations

	Magnetic Sector SIMS	Quadrupole SIMS	ToF-SIMS
Main Features			
m/Δm	High	350	High
Mass detection	Single/few masses	Single/several masses	Full range (∼10 kDa)
Primary beams	O_2^+, Cs^+ (inert)	O_2^+, Cs^+ (inert)	Ga^+ (O_2^+, Cs^+, inert)
Optimized for	Depth profiling	Depth profiling	Surface analysis
Application			
Bulk dopant	Yes	Some	N/A
Deep implants	Yes	Most	N/A
Shallow implants	Only latest tools	Yes (modern tools)	Possible (multiple guns)
Thin films	Only latest tools	Yes (modern tools)	Possible
Surface metals	Yes	No	Yes
Organics	No (elements only)	Marginal	Yes
Small-area	Several microns	Several microns	100 nm (Ga^+)
Insulators	Positive ions difficult	Yes	Yes

focused or relatively large primary ion beams, utilizing the direct imaging capabilities of the spectrometer (microscope). Most ToF-SIMS instruments are equipped with Ga^+ liquid metal ion guns, which can provide 100 nm spatial resolution.

Depth profile analysis of deep and ultrathin structures is mostly performed using either magnetic sector or quadrupole SIMS instruments. Surface sensitivity to organic and inorganic contaminants is the real strength of ToF-SIMS, because it automatically provides a survey of practically all contaminants in one analysis. When equipped with additional appropriate sputter guns, these instruments can also be used for depth profile analysis of thin films or shallow implants.

The interaction of the primary ion beam with the surface and the secondary ion formation is rather complicated and generally requires thorough investigation of potential analytical artifacts, such as secondary ion yield changes near the surface, deeper interfaces, sputter rate variations, surface roughening, ion-beam-induced dopant diffusion, etc. [203,204,218–224]. Because of potential sputter artifacts and order-of-magnitude spreads in secondary ion production for different elements in various material matrices, it is necessary to use a set of carefully prepared reference standards to calibrate the measurement. Ideally, these standards should be as close to the unknown sample as possible, both in dopant or impurity level, and in matrix composition to carry out quantitative analysis. In the optimum case, precisions near 0.5% can be realized.

28.3.2.2 Optical Spectroscopy

Purpose. The primary optical techniques for high-sensitivity detection of dopants and impurities in semiconductor solids are Fourier transform infrared (FTIR) and photoluminescence (PL) spectroscopies. These reach parts per billion atomic and lower when performed at temperatures near 4–15 K. The FTIR works best for bulk silicon, which is transparent in the infrared, and is frequently applied even at room temperature for the quantification of oxygen and carbon in Czochralski (CZ) silicon crystals and wafers. The PL at visible light frequencies penetrates only about 0.5–5 μm into silicon and hence is suitable for the analysis of epitaxial films. In some applications, low-temperature PL reaches sub-parts per trillion atomic detectability. A comparison of the methods is given in Table 28.7.

Methods. Infrared spectroscopy is based on the fact that molecules have discrete absorption frequencies associated with their vibrational and rotational motions. When a sample is placed in infrared light, it will selectively absorb at those resonant frequencies of the dopant or impurity species and pass the remainder. The absorption is associated with a change in the dipole moment of the molecule.

The transmittance (T_S, expressed in wave numbers) for a normal incidence infrared probe on a double-side-polished silicon slab of thickness d is given by

TABLE 28.7 Comparison of Optical Spectroscopy Techniques for Silicon Analysis

Technique/Reference	Sample Interaction	Application
Fourier Transform Infrared (FTIR)		
Room temperature [225–228]	Infrared absorption	Oxygen in czochralski silicon quantified at ppma level
Low temperature [229–231]		parts per billion atomic sensitivity to dopants and impurities in Si
Photoluminescence		
Room temperature [229,232]	Light emission	Spatial signal mapping of compound semiconductors
Low temperature [229,233–238]		Sub-ppb sensitivity to dopants and electrically active impurities

$$T_S = [(1-R)^2 \exp(-\alpha d)]/[1 - R^2 \exp(-2\alpha d)] \tag{28.47}$$

where R is the internal surface reflectivity and α the absorption coefficient.

For determination of interstitial oxygen in CZ silicon, for example, the 1107 cm^{-1} band related to stretching modes of Si–O bonds is measured [225,226,231]. Infrared transmission (T_F) through a float zone slab of the same thickness d (as before) combined in ratio with Equation 28.27 yields

$$T = T_S/T_F \tilde{} \exp[-(\alpha_S - \alpha_F)d] = \exp(-\alpha_0 d) \tag{28.48}$$

This is a form of the Lambert–Beer law, which relates transmitted intensity to concentration of an absorbing species. The simplicity facilitates determination of α_0, which is then converted to ppma or atoms per cubic centimeter through a constant of calibration. Calibration factors at room temperature for oxygen in silicon are available through ASTM, JEIDA, DIN, and others [225,239–243].

The FTIR is a form of infrared spectroscopy in which the absorption spectrum is acquired using an interferometer instead of a diffraction grating, so that a large acceptance aperture can be realized for increased sensitivity. The absorption spectrum is obtained by taking the Fourier transform of the measured interferogram, usually near liquid helium temperature for best performance. Spectra of dopants in silicon are typically recorded within the 200–400 wave number (cm^{-1}) range.

The PL technique floods the sample with visible light (from an Ar 514.5 nm laser, for example) to create a population inversion of electronic excited states, and relaxation through radiative emission then follows. Electrically active impurity and defect centers can be identified with ultra-sensitivity by analyzing this emission, which is also in the visible light range. Trace levels of P, B, and As can be quantitatively measured at 10^9–10^{15} atoms per cubic centimeter levels in silicon. Like FTIR, modern PL spectrometers use an interferometer for the Fourier transform photoluminescence (FTPL), which has the advantage of large aperture and increased signal intensity.

Strengths and Weaknesses. Low-temperature implementation of FTIR and FTPL is an inconvenience, but it has a sizeable sensitivity advantage for the dopant and impurity applications listed in Table 28.8.

TABLE 28.8 Strengths and Weaknesses of Low Temperature FTIR and Fourier Transform Photoluminescence (FTPL)

Measurement	FTIR	FTPL
Substitutional carbon	Yes	No
Interstitial oxygen	Yes	No
Dopants $>10^{15}$ atoms per cubic centimeter	Yes	No
Good sample throughput	Yes	No
Microspot low-temperature analysis	No	No
Dopants $<10^{12}$ atoms per cubic centimeter	No	Yes
Surface and epi analysis	No	Yes
Dislocations	No	Yes
Other deep-level centers	No	Yes

The FTIR is more suitable for bulk silicon analysis, while FTPL is more surface sensitive (0.5–5.0 μm penetration, depending on wavelength). In principle, both are non-destructive, although wafer stages for low-temperature operation are not available, so wafer samples must be broken into smaller pieces for analysis. Microspot FTIR or FTPL instruments that operate at low temperature are not yet commercially available for trace element analysis.

28.3.2.3 Radiochemical Methods for Trace Elements

Purpose. The radiochemical techniques most commonly applied to semiconductor materials are listed in Table 28.9. Of these, neutron activation analysis (NAA) is the most sensitive for trace element analysis. It is predominantly applied to bulk silicon (CZ and float zone) and can provide ppb or better sensitivities for many impurities of importance to semiconductor manufacturers. The other activation methods are more suited for profiling thin film structures.

Method. In NAA, the sample is exposed to a flux of thermal neutrons in a nuclear reactor. Radioactive isotopes are formed from stable elements in the matrix so that trace impurities can be identified by detecting gamma rays emitted from the products of (n, γ) nuclear reactions. This is normally performed for daughters of the elements irradiated that have half-lives of 24 h or more, which permits offline detection using doped Si or Ge crystal γ-detectors. For elements with daughter half-lives less than this (^{2}H, ^{11}B, ^{15}N, ^{19}F, and others), the experiment is performed at the reactor site while the sample is being neutron irradiated, and is called *prompt gamma activation analysis*.

In *charged particle activation analysis*, the sample is activated with an ion beam of suitable type and energy. Artificial radioisotopes are created by nuclear reactions on matrix atoms as well as the impurities of interest. Helium is a common projectile that offers flexibility for light elements including C, N, O, F, and Be [256]. Quantitative analysis is based on the identification of the recreated radioisotopes, most often using high-resolution γ-ray spectroscopy. Helium irradiation of oxygen in silicon, for example [^{16}O(^{3}He, p)^{18}F], produces activated ^{18}F, which is a metastable positron emitter that leads to 511 keV γ-rays characteristic of positron–electron annihilation.

In *nuclear reaction analysis*, γ-rays generated directly from the parent species are detected. The γ-energy identifies the species, while intensity relates to concentration. Hydrogen can be quantified in thin films using a ^{15}N projectile, for example. The threshold, or Q resonance, of the ^{1}H(^{15}N,α,γ)^{12}C reaction is very sharp, and occurs only at that depth in the sample where the ^{15}N projectile has slowed in the solid to exactly 6.385 ($\pm$0.005) MeV. Hence, it is possible to determine a depth profile of ^{1}H by ramping the projectile energy [252]. *Neutron depth profiling* likewise relies on the absorption of a particle (α-particle in this case) to generate depth profiles of B, N, and Li [253].

Strengths and Weaknesses. Activation techniques are inherently isotope specific and quantitative once simple geometrical calibrations of the equipment are made. They are well suited for absolute isotope quantification and generation of reference standards for other techniques. As the measurements are dependent on nuclear properties, they are completely decoupled from the chemical bonding environment.

TABLE 28.9 Radiochemical Techniques Used in Semiconductor Applications

Technique/Reference	Distinguishing Feature	Sensitivity	Application
Neutron activation analysis (NAA) [244–246]	Bulk analysis	ppta	Silicon impurities
Prompt gamma activation analysis (PGAA) [247,248]	NAA with short half-lives	ppma	Light elements
Charged particle activation analysis (CPAA) [249,250]	Particle-induced nuclear reactions	ppma	Oxygen in silicon
Nuclear reaction analysis (NRA) [251,252]	Particle-induced gamma-rays	0.1%	H profiles in films
Neutron depth profiling (NDP) [253–255]	α-Particle path length absorption	0.1%	B, Li, N profiles

Once activated, there is no chance of cross contamination with mobile ions, like Na or K. These techniques offer exceptional sensitivity for dopants and impurities and are well suited to isotope doping experiments; however, they cannot be applied with equal success to all elements of the periodic table. A reactor or accelerator facility is required, so the methods are not necessarily routinely available or low in cost.

28.3.3 Surface and Thin Film Composition and Chemistry

28.3.3.1 Rutherford Backscattering Spectrometry and Related Techniques

Purpose. Rutherford backscattering spectrometry (RBS) is a technique based in classical physics involving scattering of ionized particles by nuclei in the sample being analyzed. Common uses of RBS include quantitative depth profiling, areal concentration measurements (atoms per square centimeter), and crystal quality and impurity lattice site analysis. Its primary application is quantitative depth profiling to determine elemental compositions of thin films and multilayer structures. It is well suited for the analysis of complex silicides and multilayer metallizations, and provides data accurate to 5% or better without the use of standards. The related techniques outlined in Table 28.10 extend the methodology to ultrathin films, physical defects, impurity analysis, and hydrogen quantification.

Method. In conventional RBS, high-energy (1–3 MeV) mono-energetic ions of mass m, atomic number Z_1, and energy E_0 are directed at the surface of a sample using a particle accelerator. ^{4}He ions (alpha particles) are typically used. A small fraction of these ions pass close enough to atomic nuclei in the material so that Coulombic forces between the two nuclei cause the lighter ion to be scattered. A schematic of this process is shown in Figure 28.32. While the nuclei do not actually collide, the process can be modeled as an elastic collision using classical physics. The energy of the ion (E) after the encounter is related to the mass of the target nucleus (M) through the kinematic factor, which expresses conservation of energy and momentum

$$K = \frac{E}{E_0} = \left(\frac{\sqrt{M^2 - m^2 \sin^2 \Theta} + m \cos \Theta}{M + m} \right)^2 \tag{28.49}$$

where E is the energy of the projectile before scattering, E_0 the energy of the projectile after scattering, m the mass of the projectile, M the mass of the target nucleus, and Θ the angle of scattering in the laboratory system.

Sensitivity is near 0.01% (5×10^{18} atoms per cubic centimeter), and is related to the differential scattering cross-section ($d\sigma/d\Omega$), defined by

TABLE 28.10 Backscattering Techniques Used in Semiconductor Applications

Technique/Reference	Distinguishing Feature	Application
Rutherford backscattering spectrometry (RBS) [257–259]	>MeV particle scattering	~1 μm film composition profiles
Heavy ion backscattering (HIBS) [259,260]	Heavy incident particle	Metal impurities at high sensitivity
Medium-energy ion scattering (MEIS) [258,261]	<MeV particle scattering	~10 nm film composition profiles
Elastic recoil detection analysis (ERDA) [262,263]	Forward knock-on scattering	Hydrogen analysis in thin films
Particle-induced x-ray emission (PIXE) [264,265]	Ion-induced x-ray detection	Elemental analysis
Ion channeling [257,266]	Beam aligned to lattice	Surface lattice damage

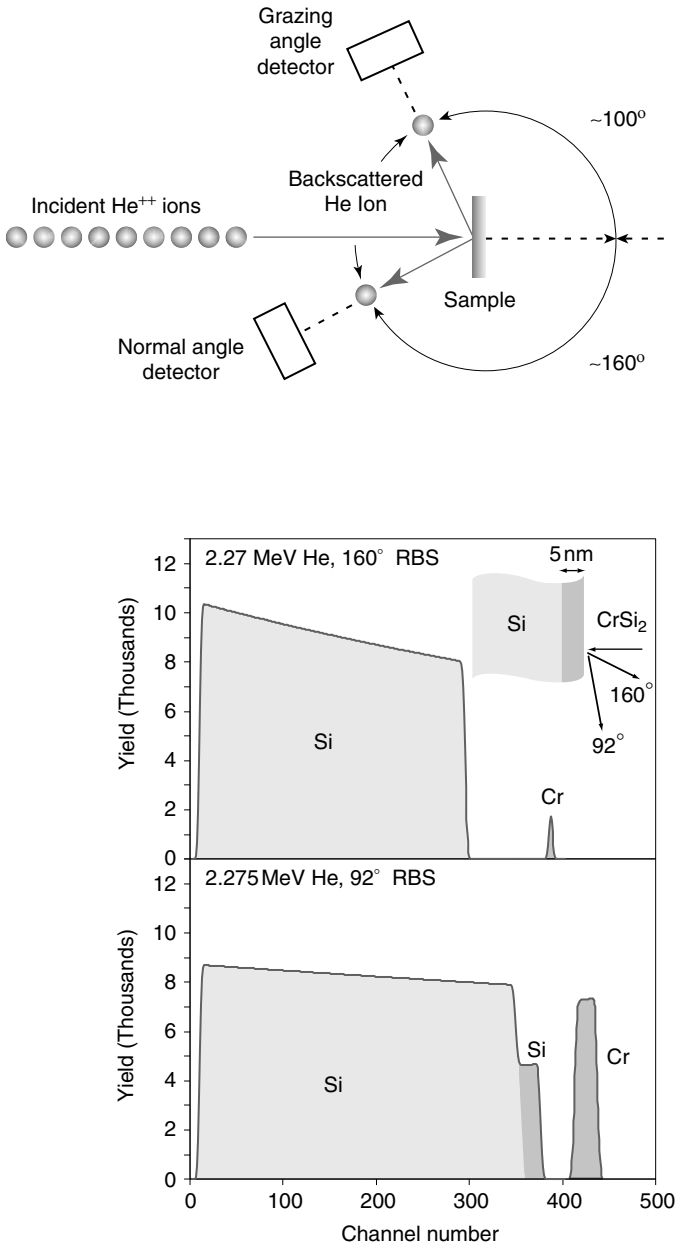


FIGURE 28.32 Illustration of the Rutherford backscattering spectrometry (RBS) process showing both normal and grazing angle configurations. The plots show how the grazing angle detector enhances surface sensitivity in the analysis of a 5 nm Cr silicide film on Si. Note how the Si signal from the CrSi film is only evident using the grazing angle detector (lower plot).

$$\frac{d\sigma}{d\Omega} = \left(\frac{Z_1 Z_2 e^2}{4E}\right)^2 \frac{4}{\sin^4 \Theta} \frac{\left(\sqrt{1 - \frac{m^2}{M^2} \sin^2 \Theta} + \cos \Theta\right)^2}{\sqrt{1 - \frac{m^2}{M^2} \sin^2 \Theta}} \tag{28.50}$$

where Z_1 is the atomic number of the projectile, Z_2 the atomic number of the target atom, and e the electronic charge.

The kinematic energy loss provides the mass of the target nucleus. Additionally, energy loss occurs as the ion slows down in the solid (dE/dx). This allows calculation of the depth of a feature beneath the surface of the sample. The scattering geometry also affects depth resolution and it is possible to improve depth resolution for thin films using a grazing angle detector (Figure 28.32). In all cases, the mass and depth scales must be deconvolved to extract the complete profile. An example of RBS data from a single thin film on substrate is shown in Figure 28.32.

In addition to elemental compositional information, RBS can also be used to study the structure of single-crystal samples. When a sample is channeled, the rows of atoms in the lattice are aligned parallel to the incident He ion beam. The bombarding He will backscatter from the first few monolayers of material at the same rate as a non-aligned sample, but backscattering from buried atoms in the lattice will be drastically reduced since these atoms are shielded from the incident He atoms by the atoms in the surface layers. For example, the backscattering signal from a single-crystal Si sample which is in channeling alignment along the ⟨100⟩ axis will be approximately 3% of the backscattering signal from a non-aligned crystal, or amorphous or polycrystalline Si. By measuring the reduction in backscattering when a sample is channeled, it is possible to quantitatively measure and profile the crystal perfection of a sample [267].

Channeling can also be used for background reduction to help improve the RBS sensitivity for light elements. For example, it is difficult to accurately measure N concentrations in TiN films deposited on Si substrates due to the overlapped signal from the Si substrate. By channeling the substrate, the substrate signal is reduced, thus improving the sensitivity for the N peak, which is superimposed on the Si signal. Since TiN layers are typically polycrystalline, the channeling does not affect the backscattering signals from the Ti or N [267].

Some of the related techniques listed in Table 28.10 are fundamentally the same as RBS. For example, heavy ion backscattering (HIBS) simply uses heavier ions for the primary beam. Collision cross-sections are higher for heavier primary ions resulting in improved sensitivities. The HIBS is particularly useful for detecting trace levels of heavy metals in light element matrices. If the incident beam has higher mass than the matrix, then the matrix elements will be forward scattered and not contribute to the signal allowing interference-free signals. Medium-energy ion scattering (MEIS) utilizes a primary ion beam in the range of 50–300 keV, which is optimum for providing classical Rutherford scattering coupled with very good surface specificity, approximately 2–10 nm. The MEIS therefore is a powerful tool for characterizing ultrathin films, shallow ion implants, trace element analysis, and studies of thin film crystallinity via ion channeling.

Elastic recoil detection (ERD) analysis takes advantage of the fact that elements in a material lighter than the primary ions will be forward scattered upon collision with the primaries. In conventional RBS, which uses ^{4}He ions, a special case of ERD known as hydrogen forward scattering is often employed to quantify the hydrogen content of thin films.

Particle-induced x-ray emission (PIXE) is an elemental analysis technique that detects x-rays that are induced by the collision of the primary particles with the atoms in the sample. The interaction causes the removal of core electrons leading to the emission of x-rays with specific energies when outer shell electrons drop to fill the core shell vacancies. The x-ray energies are independent of the excitation process and are element specific. Since these x-rays are produced constantly during an RBS analysis, PIXE requires only that an RBS instrument be fitted with a suitable x-ray detector. As an accessory on RBS instruments, PIXE is useful for heavy element identification when the elements of interest have only small differences in RBS energies but distinct differences in PIXE spectra. There are also dedicated PIXE instruments; however, these typically use H^+ bombardment instead of the He^+ used in RBS.

Strengths and Weaknesses. Since the kinematic factor and energy loss (dE/dx) curves (also known as "stopping power") are known prior to the analysis, backscatter data provide mass and depth information to about 5% accuracy or better without the use of standards. Because of the standardless quantitative analysis capability, RBS is often used to standardize results from other techniques that rely on sensitivity

factors that can vary with the sample matrix, such as Auger. The technique works best for heavy elements in a light matrix, but can be extended to thin oxynitride thin films, for example, with the medium-energy modification. Backscattering requires an accelerator capable of producing 1–3 MeV energy H or He ion beams within a 1 mm spot on the sample. Hence, RBS is not a small spot technique, but with specialized high-energy ion optics, it is possible to reduce the spot size to 2 μm [268].

28.3.3.2 X-Ray Fluorescence and Total Reflection X-Ray Fluorescence

Purpose. X-ray fluorescence (XRF) is ideal for rapid qualitative and quantitative analysis of atomic constituents. It is a particularly useful tool for the initial analysis of an unknown contamination, in that it accommodates solid or liquid samples, metals, and insulators, and production size wafers. With proper calibration, it can be used to monitor the thickness of metal films between 20 nm and several microns thick, with precisions as high as 0.01% [269–271]. The XRF is basically a bulk evaluation method, with a sampling depth in the 10 μm range, determined by penetration of the incident x-ray radiation and the escape depth of the characteristic fluorescence. Detection limits for XRF are on the order of 10 ppm.

Total reflection XRF (TXRF) is a special case of XRF for which dedicated instruments are available. In TXRF, the x-rays impinge on the sample at a grazing angle, less than the critical angle (ϕ_c) for total reflection (in the mrad range). In this manner, the excitation depth is limited to a few nanometers, giving rise to a strong fluorescence signal from the near surface of the sample. The TXRF requires the target sample to have a smooth surface and is commonly used in the semiconductor industry for measuring transition metal contamination on wafer surfaces, either single-point measurements or complete maps of entire wafer surfaces [272–274]. For semiconductor applications, TXRF instruments, which can be fully automated, typically have at least the sample introduction area housed in a cleanroom environment to avoid contaminating sample surfaces with airborne particulate matter prior to analysis.

Method. In XRF, incident x-rays are used to eject inner shell electrons from atoms of the sample via the photoelectric effect. The atom then relaxes through the emission of an x-ray with energy characteristic of the parent atom and intensity proportional to the amount of the element present. Conventional instruments use a stationary x-ray target, typically made of elemental Cr, Cu, Ag, Mo, W, or Au (with characteristic energies of 5.411, 8.041, 22.104, 17.443, 9.60, or 9.711 keV, respectively). The spectrum of x-rays that irradiate the sample include the broad band Bremsstrahlung background. X-ray detectors may be of the EDS or wavelength dispersive spectroscopy (WDS) type. Calibration for quantitative measurement of film thickness requires carefully prepared reference samples of identical composition.

A schematic showing the configuration of a TXRF instrument is shown in Figure 28.33. The TXRF requires a specular surface since the primary angle of incidence is extremely small ($< \phi_c$), and is ideal for highly polished samples like silicon wafers. The critical angle (in mrad) is given by

$$\varphi_c = 3.72 \times 10^{-11} \sqrt{\frac{n_e}{E}} \tag{28.51}$$

where ϕ_c is the critical angle, n_e the electron density (in cm^{-3}), and E the x-ray energy in keV [272]. For a typical $W_{L\beta}$ source (9.66 keV) on a silicon wafer, ϕ_c is about 3 mrad (0.2°).

When the incident x-ray beam hits a smooth surface at angles less than ϕ_c, total *external* reflection occurs. Under these conditions, where the incident beam is 100% reflected, an evanescent wave is formed at the reflecting surface. The penetration depth of this wave is defined as the depth where the intensity decays to $1/e$ (37%) of its initial value, generally limited to a few nanometers. It is this evanescent wave that generates the fluorescent signal within the solid, hence resulting is a strongly surface sensitive measurement.

The sensitivity of TXRF is in the 10^9–10^{12} atoms per square centimeter range, although improvements in instrument configuration and specialized synchrotron applications are pressing this detection limit range several orders of magnitude lower.

A technique that can be used to maximize TXRF detection sensitivities is the vapor phase decomposition (VPD), which concentrates the impurities by chemical evaporation before the analysis

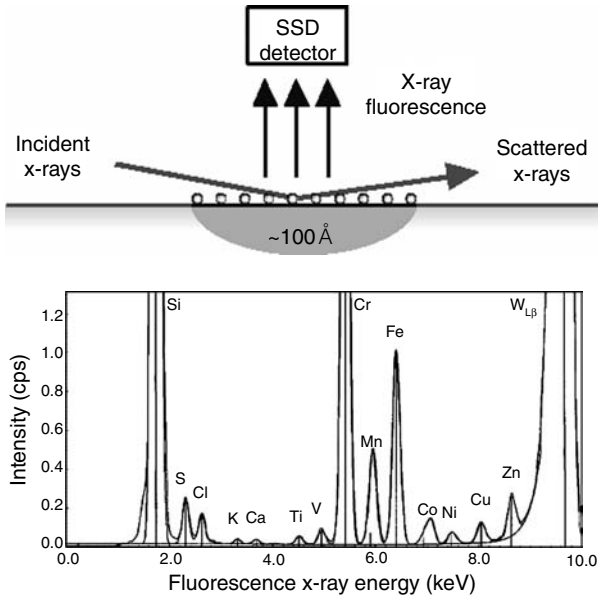


FIGURE 28.33 The upper panel shows the configuration for a total reflection x-ray fluorescence (TXRF) measurement. The lower panel shows a typical TXRF spectrum from a Si wafer analysis.

is performed. In this procedure, a wafer is exposed to a dilute HF vapor, which slowly etches any silicon oxide on the surface of the wafer, subject to reactions of the type

$$6HF + SiO_2 \rightarrow H_2SiF_6 + 2H_2O \text{ (residue)} \tag{28.52}$$

In effect, this integrates over the etched depth by removing silicon matrix atoms and leaving the impurity behind. Variations of the VPD technique include fluid drop scanning, in which a small droplet of HF is moved systematically over the surface of the wafer by a custom-designed holder that keeps it from slipping away. The drop, containing the collected residue, is then left to dry on the wafer for TXRF analysis.

Strengths and Weaknesses. Conventional XRF provides rapid identification for samples of unknown composition or origin, either in solid or liquid form. Elemental qualitative composition to 5–10 wt% accuracy can be determined in a matter of 15–20 min. Accuracies approaching 0.01 wt% are possible, provided suitable reference samples of similar composition and uniformity are available. First principles determination of wt% based on raw x-ray intensities is more difficult with XRF, because the interplay between the x-ray energy, matrix absorption, and secondary fluorescence introduces significant complications [275].

Total reflection XRF has become an industry standard for quantifying trace metal impurities on semiconductor wafers (Figure 28.33). Instruments are available, which are fully automated, such that wafer boxes can be introduced directly into the instrument, requiring no human contact with individual wafers. Outside of fabs, TXRF instruments can have sample introduction and removal stations housed in cleanroom environments to avoid particulate contamination of the wafer surfaces. Its strengths include simultaneous multi-element analysis with very low detection limits and simple quantification by internal standardization over a wide dynamic range. In addition, it is a non-destructive measurement technique [274].

Generally, TXRF is not suitable for the detection of low-Z elements ($Z < 14$), attributable to problems associated with fluorescence excitation, energy-dispersive detection, and quantitative analysis [274]. The TXRF is also sensitive to surface roughness, which increases the background signal and degrades the detection limits. If the roughness is great enough, the conditions necessary for total reflection will not be met and the result will be a glancing angle XRF measurement instead of a TXRF measurement, invalidating the TXRF quantification procedures.

Since it is difficult to focus x-rays, small spot XRF tools are not readily available for high spatial resolution. Most commercial instruments provide a 100–500 μm spot, with 30 μm reported in the extreme [276]. Recent synchrotron experiments approach 0.5 μm [277], and capillary optics may offer additional improvement [278]. In commercial TXRF instruments, the spot size is determined by the EDS detector area (1 cm^2). The beam is intentionally made wide enough to excite the entire area under the detector in order to achieve the required detection limits (Figure 28.33).

28.3.3.3 X-Ray Photoelectron Spectroscopy

Purpose. X-ray photoelectron spectroscopy (XPS), also known as electron spectroscopy for chemical analysis (ESCA), is an analytical technique that can provide information about the surface and near-surface region of materials and, for very thin layers, interface chemistry. With sputter depth profiling, XPS can provide composition (and sometimes chemistry) as a function of depth into the sample. The XPS is sensitive to the outermost atoms or molecules of a sample (typically 0–10 nm) with sub-monolayer detectability. All elements with the exception of H and He can be detected and quantified. Since atoms such as C, N, O, and F are the major constituents of most surface contamination including airborne molecular contamination, surface corrosion or oxidation, and residues from cleaning or process steps, XPS is a valuable chemical characterization tool applied to all stages of semiconductor manufacturing [279].

X-ray photoelectron spectroscopy is also very useful for determining composition and chemistry of deposited or grown thin films. For films on the order of 10 nm thick or less, an XPS surface analysis can provide a "bulk" characterization of the film. In recent years, XPS has been routinely used for characterizing ultrathin materials such as barrier metals and gate dielectric materials. For instance, XPS can measure thickness and nitrogen dose in ultrathin silicon oxynitride films with extremely high precision [280].

Thicker films may require sputter depth profiling to provide film composition and chemistry. Sputter depth profiling is accomplished by alternating ion etch steps and data acquisition steps within the sputter crater to determine elemental concentrations as a function of depth. The sputtering conditions can be varied to provide etch step sizes of < 0.5 nm to > 1 μm per step. The XPS depth profiling is suitable for bulk thin film (single or multilayer) analysis as well as for investigating buried interface chemistry.

Method. In XPS, the sample surface is bombarded with x-rays, typically $Al_{K\alpha}$ (1486.6 eV) or $Mg_{K\alpha}$ (1253.6 eV), causing the discharge of core level electrons, a process known as the *photoelectric effect*. Ejected photoelectrons are of discrete energies relating to the specific parent atoms, given by the expression

$$\mathrm{KE} = h\nu - \mathrm{BE} + \varphi \tag{28.53}$$

where KE is the kinetic energy of the photoelectron (in eV), $\hbar$ Planck's constant, ν the frequency of incident x-ray, BE the binding energy of the electron in the atom (in eV), and ϕ the spectrometer work function ($\sim$3–4 eV).

Photoelectrons are generated within the x-ray penetration depth (typically many microns), but since the photoelectron energies are low (less than 1486 eV for $Al_{K\alpha}$), only the photoelectrons within the top three photoelectron escape depths are detected. This is the origin of surface selectivity for XPS [281]. Escape depths are on the order of 1.5–3.5 nm, which leads to an analysis depth of approximately 5–10 nm. Typically, 95% of the signal originates from within this depth.

Quantitative analysis with XPS is accomplished by determining the atom fractions of each constituent and normalizing to 100% of the detected elements. The atom fraction of constituent element C_x is

represented by the following equation:

$$C_x = \frac{(I_x/S_x)}{\sum(I_i/S_i)} \tag{28.54}$$

where I is the intensity of the photoelectron peak (measured as the peak area) and S the atomic sensitivity factor. While atomic sensitivity factors may be somewhat matrix dependent, primarily due to differences in the photoelectron mean free paths and to a lesser extent the photoelectron cross-sections in different materials, the ratios of these values are nearly constant. The result is nearly matrix-free quantification. Therefore, RSFs for a given spectrometer can be used to provide atomic percent of the detected elements with absolute accuracies on the order of 10% or better if the peaks have sufficient signal-to-noise ratios.

A unique property of XPS rests on the fact that the core level electrons nearest to the valence shells often exhibit shifts in binding energy due to the specific chemical environment of the atom. For example, consider the binding energy of Si in various chemical forms: elemental Si and some silicides have photoelectron binding energies of approximately 99 eV, Si in silicon carbide (SiC) a binding energy of about 100.6 eV, in silicon nitride (Si_3N_4) a binding energy of approximately 101.8 eV, and in SiO_2 a binding energy of approximately 103.3–103.5 eV. Figure 28.34 shows a typical nitrogen spectrum from an ultrathin silicon oxynitride gate dielectric film where multiple bonding states for nitrogen are observed. Binding energies for elements in various chemical forms have been tabulated [282].

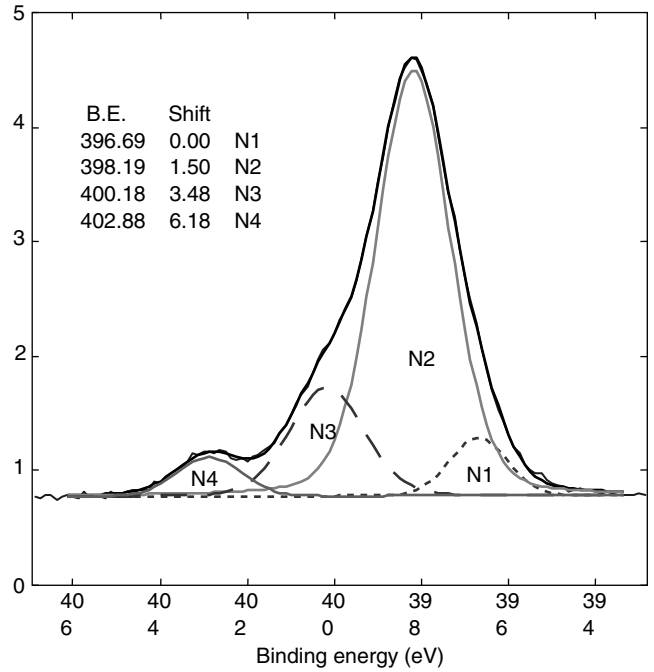


FIGURE 28.34 X-ray photoelectron spectroscopy high-resolution spectrum of nitrogen from an ultrathin silicon oxynitride gate dielectric film. Several chemical forms of N are identified, including Si_3N_4 (pure silicon nitride, peak N1) and three SiO_xN_y peaks. Peaks N2 and N3 result from bonding arrangements where one or more oxygen atoms replace nitrogen in the Si_3N_4 tetrahedral coordination. The highest binding energy peak (N4) represents the N–O bond from a nitroso or bridging bond. In this example where the peaks partially overlap, non-linear least squares (NLLS) curve fitting was applied in order to determine the relative percentages of each species within the analytical volume.

Strengths and Weaknesses. XPS is a very good survey technique for characterizing surfaces or residues in cases where the composition is unknown. The ability to provide chemical bonding information coupled with quantitative analysis also makes the technique useful for characterizing chemical changes at surfaces due to material processing. Furthermore, the sampling depth makes it feasible to obtain bulk characterization of very thin films on the order of 10 nm or less.

Perhaps one of the most useful attributes of XPS is in analyzing insulating materials. Because XPS probes the surface with electrically neutral photons and the interaction only ejects photoelectrons (resulting in a positively charged surface), it is simply a matter of replacing the ejected electrons to return the surface to charge neutrality. This is commonly done by "flooding" the area with low-energy thermal electrons that are then electrostatically attracted to the positively charged areas, thereby restoring charge neutrality.

The detection limits for XPS are on average in the part per thousand range under typical analysis conditions. Therefore, the technique is not generally suitable for trace elemental analysis. The sensitivities for some, usually, heavier elements may be nearly an order of magnitude better than the average, while the very lightest elements (Be, Li) have sensitivities closer to 1 atomic percent.

Sputter depth profiling can produce artifacts depending on the material being sputtered. Artifacts can be incorrect stoichiometries due to preferential sputtering of different components within the material, or degradation of the materials under the energetic ion bombardment such that chemical state identifications are not reliable. Recent advances in ion sources have made strides in mitigating the latter problem [283].

Scanning XPS for the purpose of imaging is not easy to achieve since it is extremely difficult to steer or raster the x-ray beam itself. However, it is possible to scan the XPS analyzer spot (within a large x-ray irradiation area) to achieve XPS elemental maps. Another approach is to use a scanned electron beam source over the x-ray anode surface to effectively produce a scanned x-ray beam. These methods, however, only provide spatial resolution on the order of a few microns or so.

28.3.3.4 Auger Electron Spectroscopy

Purpose. Auger electron spectroscopy (AES) has many similarities to XPS, including a comparable range of detection limits and sampling depth. Like XPS, AES can detect all elements except for H and He and therefore serves as a useful survey measurement for surface characterization. Sputter depth profiling provides a means for bulk characterization of thin films or film stacks as well as measuring interface composition. The depth of information for AES extends from the surface up to a maximum of approximately 10 nm depending on the energy of the measured Auger electrons. However, the majority of Auger electrons for many common elements occur at relatively low kinetic energies, which results in an average sampling depth for AES of 3–5 nm (shallower than the average XPS sampling depth) [284].

Auger electron spectroscopy uses a focused electron beam for excitation, which extends the analysis to localized spots as small as 15 nm across. The small spot capability makes AES suitable for analyses such as characterization of via opening cleanups, small particles, or general surface contamination confined to areas too small for XPS analysis. Auger depth profiles are common in the semiconductor industry because of its ability to obtain profiles from small areas. The AES is often used in conjunction with FIB to obtain elemental compositions of buried defects or layers. The FIB is used to prepare cross-sections which are then analyzed by AES. There are AES instruments available that integrate a FIB column for this purpose.

Method. In AES, the sample is bombarded with a focused electron beam causing the ejection of an inner shell electron, similar to XPS. The Auger process, first described by P. Auger [285], is an electronic rearrangement that serves to relax the atom from the excited state brought on by the inner shell vacancy. It is a two-electron process first involving an electron from a higher energy shell filling the inner shell vacancy followed by a second electron from the higher energy shell leaving the atom for energy conservation. It is this latter *Auger electron* that provides the basis for AES. Electrons resulting from the Auger process have discrete energies relating to the atomic number of the parent atom. As with photoelectrons, the Auger electrons are low-kinetic-energy electrons and the associated mean free paths ensure that only the Auger electrons from near the surface escape without energy loss, hence providing the surface sensitivity of the technique.

Quantitative analysis of Auger data is very similar to that for XPS. Most algorithms for concentration estimates involve the incorporation of intensity ratio measurements into equations of the type

$$P_i = \frac{(I_i/S_i)}{\sum_j (I_j/S_j)} \tag{28.55}$$

which express the atomic percent (P_i) of element i as a function of the total Auger current I_i, sensitivity factor S_i, and corresponding ratios for all of the other elements detected [119,120,286]. Because the electron beam probe–sample interaction also produces secondary and BSEs that contribute to a generally increasing background with increasing energy, the Auger electron peaks situated on top of this background can be rather small and, historically, were more difficult to detect. For this reason, peak-to-peak values from differentiated Auger spectra ($\mathrm{d}(E \times N(E))/\mathrm{d}E$) are usually used in place of I_i. However, this practice has changed somewhat in recent years. Quantitative algorithms now rely on the measurement of peak areas taken directly from peaks in the $N \times N(E)$ spectrum [287].

Sensitivity factors have been adapted to achieve the desired balance between quantitative accuracy and simplicity [288]. The most convenient and widely applied method is to use published values for pure element standards. Errors in this approach can be quite large (20%–50%), but they are often not cause for concern in production applications where gross problems can be resolved with semiquantitative estimates. In cases of heavily contaminated surfaces, errors in excess of 80% have been willingly tolerated [289]. However, this method of using pure element sensitivity factors can give 2%–5% reproducibility in sequential measurements.

Strengths and Weaknesses. The similarities between AES and XPS mean that the techniques share some of the same strengths and weaknesses. The AES is a good survey technique rendering itself useful for characterizing unknown contaminants or materials. The AES is not considered a trace analysis technique. The small spot capability makes Auger the best technique for determining compositions of submicron diameter particles, as well as surface or depth profile analysis in small areas. The superior imaging capabilities of modern AES instruments combined with the speed at which a typical AES survey spectrum can be acquired is on par with most SEM/energy dispersive X-ray systems. Coupling AES with FIB cross-sectioning makes the technique a powerful tool for characterizing buried defects or layers (Figure 28.35).

In practice, modern AES instruments benefit from the use of a rastered electron beam to provide real-time SE imaging of the analysis areas/defects (referred to as scanning Auger microscopy or SAM). Therefore, in addition to Auger survey spectra, SE images, BSE images, Auger elemental maps, and Auger elemental line scans are typical supporting data available in a successful AES analysis.

Because the excitation source is electron bombardment, AES is very susceptible to charging and is generally not suitable for insulating materials. However, thin insulating layers on a conductive substrate can often be accommodated if the primary electron current penetrates into the underlying conductive layer.

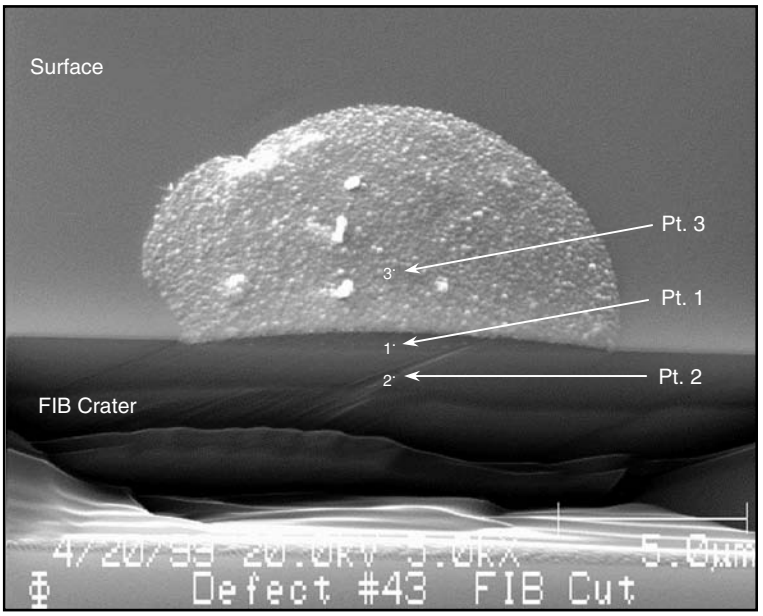
The mechanics of depth profiling by AES is identical to that of XPS. Depth profiles are typically prepared by alternating ion etching steps with the data acquisition steps. As with XPS, artifacts are introduced if the exposed surface becomes rough or damaged [218,290,291].

28.3.4 Stress and Physical Defects

28.3.4.1 X-Ray Diffraction

Purpose. XRD is able to unambiguously identify the composition of crystalline and polycrystalline semiconductor powders, thin films, and substrates. It also provides unique information about the strain, grain size, crystalline phases, preferred orientation, and defect structure of polysilicon Al, Cu, Au, and other metal layers.

Method. Diffraction techniques require that the sample contain single-crystal or polycrystalline components. On the nanometer scale, these are made up of parallel planes that are spaced at regular intervals, given by the cell constants a, b, and c. For the familiar cubic crystal, these form an orthogonal



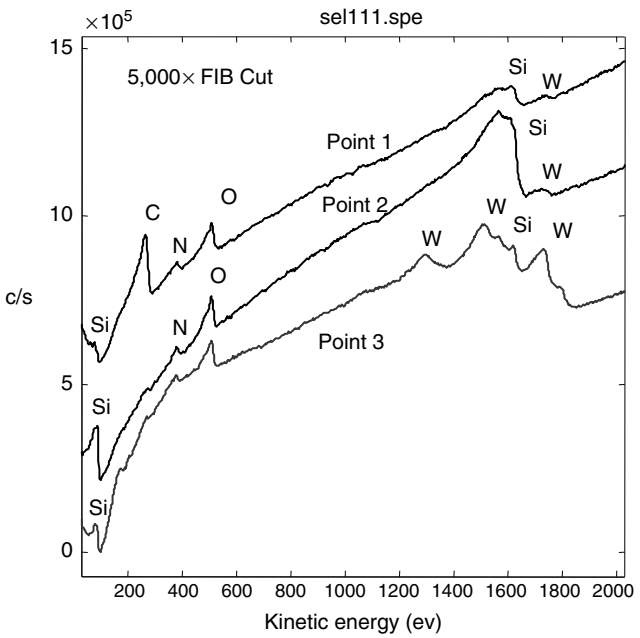


FIGURE 28.35 The FIB cross-section of a buried defect on a wafer surface after WSi deposition. The Auger electron spectroscopy analysis from the surface (point 3), center of cross-sectioned defect (point 1), and sectioned surface below defect (point 2) indicate that the defect is carbonaceous.

TABLE 28.11 Diffraction Techniques Used for Semiconductor Applications

Technique	Information Provided	Reference
Kinematic Theory		
Powder diffraction[a]	Lattice parameter for identification	[293,294]
Laue diffraction[a]	Crystal orientation	[293,295]
Seeman–Bohlin diffraction[a]	Crystalline phases in thin films	[293,294]
Pole figures[a]	Crystallographic texture and orientation	[135,296]
High-precision cell constants[a]	Quantitative strain analysis	[297,298]
Bragg spacing comparator	Oxygen effect on Si lattice parameter	[298,299]
Dynamic theory		
Anomalous x-ray transmission	Gauge of crystal perfection	[300,301]
Diffuse x-ray scattering	Bulk precipitate size and number density	[299,302]
X-ray reflectivity[a]	Nanometer roughness and defects	[303,304]

[a] More frequently used.

coordinate system. An arbitrary set of planes can be described by the Miller indices (h, k, l), which define where the plane intersects the coordinate axes in integer multiples of a, b, and c, respectively. The distance (d) between (hkl) planes in a cubic crystal ($a=b=c$) is given by

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \tag{28.56}$$

The condition for constructive interference of x-rays of wavelength λ from planes of spacing d_{hkl} is given by Bragg's law,

$$n\lambda = 2d_{hkl} \sin \theta_{hkl} \tag{28.57}$$

where θ_{hkl} is the angle between the atomic planes and the incident (and diffracted) beam, and n is a positive integer (1, 2, 3, …) denoting the order of the reflection.

The diffraction techniques listed first in Table 28.11 depend on this equation, which is referred to as the *kinematic* description. Those listed last are more appropriately described by the *dynamical* theory, in which the interactions of x-ray wavelets at all points throughout the irradiated volume are taken into account. The examples in Figure 28.36 illustrate a practical case of using both XRD and x-ray reflection for characterizing a high-k dielectric film.

Procedures and tools for evaluation of reference x-ray powder patterns rely on the powder diffraction file (PDF) published by the Joint Committee on Powder Diffraction Standards (JCPDS) of the International Centre for Diffraction Data [292,293]. These are in excess of 30,000 reference spectra relevant to semiconductor applications and are available on CD-ROM or as database upgrades for microcomputer search and retrieval.

In residual stress measurements by diffraction, the strain in the crystal lattice is measured, and residual stress then calculated assuming linear elastic distortion of the crystal lattice. Although the term *stress measurement* has come into common use, stress is an extrinsic property that is not directly measurable. High-precision changes in cell constants based on Bragg's law can be performed with a four-circle diffractometer to determine the strain. Values of $\Delta d/d$ as small as 10^{-6} are possible, and in special experiments (Bragg spacing comparator) this can be extended to 10^{-9}. Also, the topography methods covered in the following sections can be applied to measure strain induced by a thin film deposited uniformly over a wafer, as can x-ray linewidth broadening.

Strengths and Weaknesses. XRD techniques are considered to be non-destructive, and suitable for contactless process monitoring, although few such applications are found today in the wafer fab environment. They are generally fast, requiring 10–30 min acquisition time for a good spectrum or lattice parameter measurement. Extensive diffraction tables are on CD-ROM (JCPDS) for rapid search

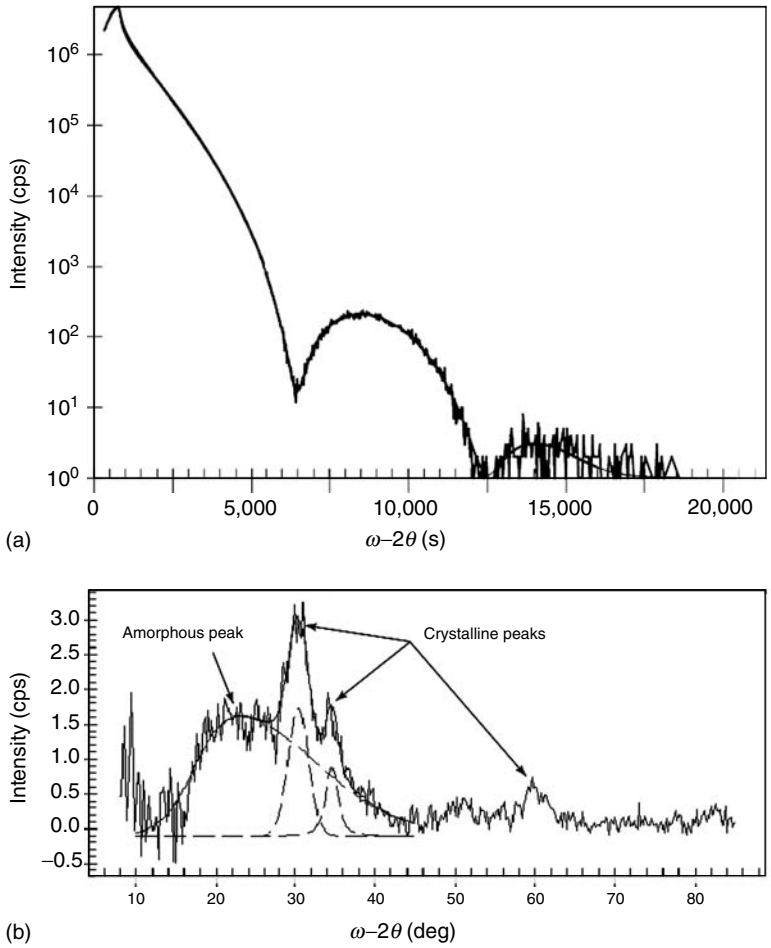


FIGURE 28.36 X-ray reflectance analysis pinpointed the thickness of a HfSiON thin film at 25.02 Å (a). By measuring the crystalline and amorphous peaks from the same film, an x-ray diffraction (XRD) analysis determined that the crystalline fraction was 51.7% (b). Data courtesy of Bede X-ray Metrology. Sample courtesy of Manuel Quevedo, SEMATECH.

and retrieval. X-ray generators of the rotating anode type provide high power (10–18 kW) relative to fixed target sources (4–5 kW), which are typically more reliable and less costly.

Small spot analysis is limited with XRD techniques in general, except in special experiments involving an intense synchrotron source, or advanced Fresnel zone plate lenses [305]. With the Gandolphi camera, it is possible to measure patterns from a single particle as small as 30 μm in diameter with a laboratory setup, but the technique may require hours of acquisition time.

28.3.4.2 X-Ray Topography/X-Ray Reflectance

Purpose. X-ray topography refers to a detailed description and mapping of physical features of a crystalline solid, such as a silicon wafer, either throughout the bulk or in the near-surface region, depending on the camera used. The images formed reveal surface relief, wafer warpage, small changes in crystallographic orientation, strains associated with epi-films and lattice defects, oxygen precipitates, and thermal process deformations.

TABLE 28.12 Topography Techniques Applied to Semiconductor Materials and Processing

Technique	Information Provided	Reference
Surface		
Berg–Barrett topography	Wafer surface defect image	[301]
Double-crystal topography	High-strain resolution images	[297,306,307]
Triple-crystal topography	High-strain resolution images	[297,299,307,308]
Bulk		
Scanning Lang topography	Wafer volume defect image; wafer warpage	[301,309]
Section topography	Wafer cross-section defect image	[299,310,311]
Moire Lang topography	Superimposed SIMOX (separation by implantation of oxygen) lattice rotation	[312]

Method. In x-ray topography, the intensity of a select Bragg diffraction spot is measured as a function of position across a crystalline solid, such as a silicon wafer. Deviations in the intensity relate to small changes in the variables of the Bragg equation 28.57 brought about by defects in the crystal lattice, or inclusions such as dopants, precipitates, and impurities in the material. *Kinematic theory* (the Bragg equation) accounts for intensities from mosaic structures such as imperfect silicon, polysilicon, or aluminum metallization, which consist of many small slightly misaligned grains. The dynamical theory of diffraction treats the stronger wave interactions present in nearly perfect single crystals, including silicon wafers, for which multiple diffraction and extinction effects occur.

A variety of topographic cameras are available for forming images of the surface or bulk of wafers, as listed in Table 28.12. The surface methods may flood the wafer with $Cu_{K\alpha}$ (8.041 keV) x-rays at grazing incidence (Berg–Barrett), for example, or position it at a steeper angle for more penetration into the solid. This may be as little as 0.1–1.0 μm, or as much as 20–150 μm, depending on the geometry. Incident x-rays of higher energy, such as $Mo_{K\alpha}$ (17.443 keV), penetrate entirely through the wafer, and are recorded on the backside by a sheet of .photographic film. These are the transmission methods, which probe the bulk using scanning (Lang) or stationary (section topography) cameras.

The lateral resolution in topography is not limited by wavelength of the x-rays, but by grain size of the recording medium, which for the best nuclear emulsion films is about 1 μm. Although this is clearly inadequate for direct imaging of point defects, dislocations, and atom aggregates, the strain fields associated with these generally extend over distances of several microns or more. This is illustrated in Figure 28.37, where transmission mode (bulk analysis) and reflection mode (surface) scans are used to map defects both in and on Si wafers.

The superior strain sensitivity of double- and triple-crystal configurations derives from the low divergence of x-rays presented to the sample by a high-quality reference crystal. The expression relating diffracted intensity contrast ($\Delta I/I$) at a Bragg angle Θ to defect-induced lattice distortions is given by

$$\Delta I/I \propto (\Delta d/d)\tan\Theta + \Delta\alpha + \Delta\beta \tag{28.58}$$

The sample descriptives are $\Delta d/d$ (relative change in lattice spacing) and the component of local lattice rotation, $\Delta\alpha$. It is evident that any offset ($\Delta\beta$) caused by divergence in the incident probe must be small, relative to these for a successful measurement to take place. With oxygen defects in silicon, for example, both $\Delta d/d$ and $\Delta\alpha$ assume a range of values between $\pm 5\times 10^{-6}$ for as-grown wafers (no precipitation annealing). This is beyond the reach of Lang topography (10^{-4}), but is detectable with the double-crystal methods (10^{-6}).

Strengths and Weaknesses. As with the diffraction techniques, x-ray topography is considered to be non-destructive, although prolonged exposure of wafers with integrated circuit patterns could in principle damage sensitive junctions and thin oxides. A high-power (10–18 kW) rotating anode x-ray generator is required for the double- and triple-crystal cameras to compensate for large intensity loss from the reference diffracting crystals, which are used to reduce divergence in the incident beam. It is also of advantage for the other methods of Table 28.12, rendering acquisition times of 45–100 min feasible in most cases.

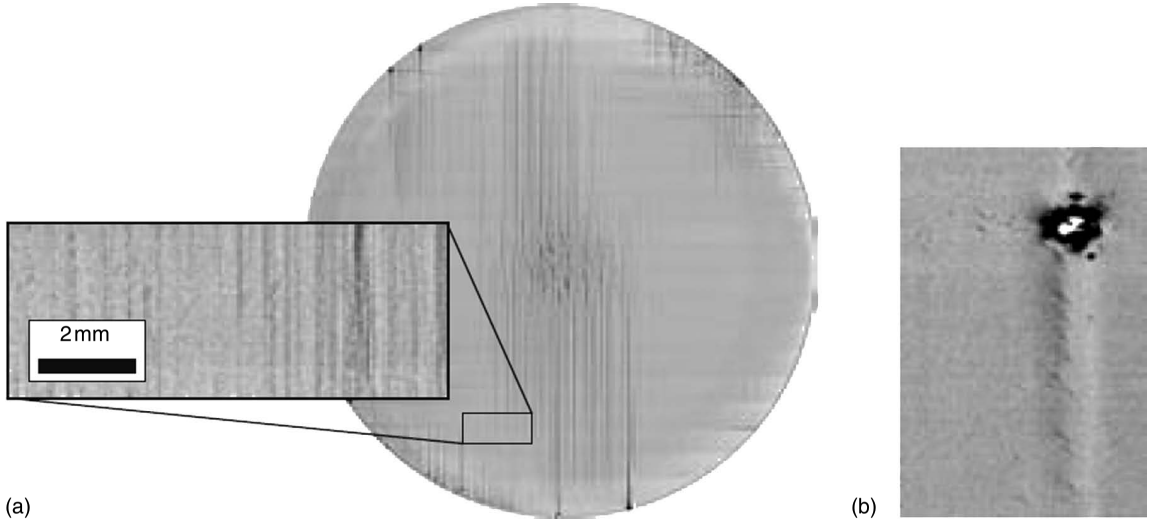


FIGURE 28.37 XRD imaging (i.e., x-ray topography) was used in a transmission mode to provide detailed images of thermal slip dislocations in a wafer (above left). The transmission mode measurement allows these slip bands to be detected at an early stage from within the bulk of the wafer. Data courtesy of Bede X-ray Metrology. XRD imaging can also be utilized in a reflection mode to image surface defects, such as the pin mark with surrounding threading dislocations (above right). The scan (at right) was obtained at a resolution of 5 μm. Data courtesy of Bede X-ray Metrology. Sample courtesy of Dr. Frans Voogt, Phillips Semiconductor, the Netherlands.

28.3.4.3 X-Ray Rocking Curves and Pole Figures

Purpose. Rocking curves provide quantification of the information available in topography images. They provide a measure of surface polish quality, strain relating to implant and thermal processing, orientation anomalies, as well as lattice match and interface integrity of heterostructure stacks based on silicon. Large collections of rocking curves (*pole figure* projection) are useful for determining preferred grain orientation in silicide and thin metal films, which relates to electromigration, voiding, and metal interconnect reliability.

Method. X-ray rocking curves exploit the detailed information available in a single Bragg diffraction peak acquired from a highly crystalline solid, such as a silicon wafer or epitaxial film [297,301]. The *rocking curve* itself is measured by recording the intensity profile of the reflection as the sample is continuously turned through a small angle (5–10 s of arc in most cases) in and out of the diffracting condition. Rocking curves are related to x-ray topography images and can be acquired using the same double- and triple-crystal cameras. However, they are generally more quantitative than topography, in that numbers can be extracted based on the width, height, and skirts of the intensity profile.

If the sample is a polycrystalline film or substrate, it is no longer suitable for rocking curve quantification. In principle, however, it is possible to acquire individual rocking curves from each grain, provided the x-ray source is focused small enough to sample one grain at a time. Such a measurement would indicate crystal quality of each grain, as well as orientation with respect to the incident probe. The collection of many such orientations is the basis of a pole figure, in which grain tilts are plotted in a spherical polar projection.

In practice, such small x-ray probes are not available in the analytical laboratory, although focused electron beams have been used in a similar way to map grain orientations within patterned metallizations [135]. Generally, broader beams are applied to large aggregates of grains embodied within a continuous film of polycrystalline silicon or aluminum, for example. In this case, the entire sample is rotated about small angles beneath the incident x-rays, using a specially constructed pole figure camera. As each grain comes into the diffracting condition for a given reflection, one or more points of intensity are recorded on the projection at positions commensurate with its tilt away from the surface normal. Pole figures provide an average measure of preferred grain orientation.

Strengths and Weaknesses. As with the topographies, a high-intensity rotating anode x-ray generator (10–18 kW) is preferable for rocking curve acquisitions to compensate for reference crystal intensity loss, although a weaker fixed target source (4–5 kW), can be used when high sample throughput is not as important. Detailed analysis of the fine structure in rocking curve profiles is based on theoretical models, which rely on iterative numerical computation for convergence. However, these are able to provide unique strain profiles relating to implant, anneal, and etch processes. X-ray techniques require special attention to safety procedures, which can be managed effectively.

28.3.4.4 Raman Spectroscopy

Purpose. Raman Spectroscopy is an analytical technique that utilizes scattering of laser light from materials in order to measure vibrational frequencies of chemical bonds. While Raman has many capabilities for material characterization of solids, liquids, or with proper instrumentation, gases, a unique ability of the technique is measuring stress or strain in materials or thin films. The magnitude and the nature of stress/strain (tensile or compressive) can be evaluated by comparing the observed frequencies of Raman bands from strained materials with those of strain-free references (Figure 28.38). The peak shifting is approximately linearly proportional to the magnitude of the strain. Some examples of materials where strain has been measured by Raman spectroscopy include [313]

1. semiconductor heteroepitaxial layers and strained superlattices,
2. heterostructures consisting of elements with different thermal expansion coefficients, such as SOI structures, and
3. semiconductor surfaces prepared by polishing, ion implantation, ion etching, etc.

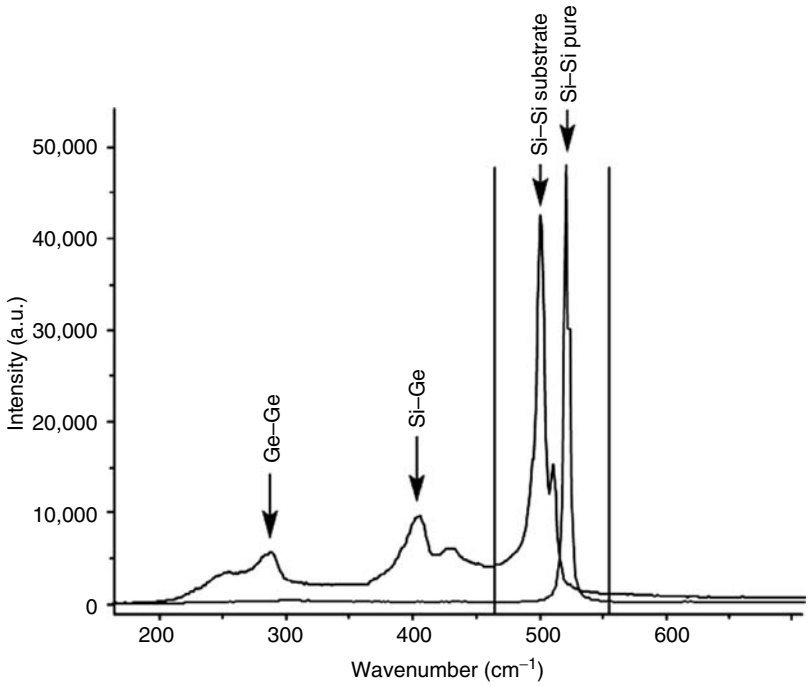


FIGURE 28.38 Raman spectrum obtained from a 0.1–0.2 μm thick Si epilayer grown on an SiGe substrate with 30% Ge in it. The Raman spectrum has spectral contribution from both the epilayer and the substrate. The Si–Si epilayer phonon vibration appears as a shoulder at 510.9 cm^{-1} on the stronger Si–Si substrate vibration at 499.9 cm^{-1}. An overlaid spectrum from a stress-free Si sample shows that the Si epilayer vibration is red shifted by ~9.5 cm^{-1} from that of the pure Si, demonstrating a tensile strain in the Si epilayer. The cause of the strain is the difference in lattice parameters between Ge and Si.

Method. The Raman effect results from the interaction of vibrational motions of molecules with electromagnetic radiation. In Raman spectroscopy, a sample is irradiated by an intense laser beam in the UV–visible range of frequency ν_0. The light scattering process involves both Rayleigh and Raman scattering. Rayleigh scattering accounts for $>99.9\%$ of the scattered light and has the same frequency as the incident beam (ν_0). The significantly weaker Raman scattered light has frequency $\nu_0 \pm \nu_m$, where ν_m is a vibrational frequency of a molecule from the sample. The two components of Raman scattering are known as the *Stokes* and *anti-Stokes* lines, the Stokes lines being the stronger of the two. By measuring the shifts in frequency of the Stokes (and anti-Stokes) lines from the incident beam frequency, a vibrational spectrum characteristic of the molecular bonds making up the irradiated sample can be constructed [313,314].

Commercial Raman microscopes can achieve lateral resolutions of approximately 1 μm. Raman microscopes can also be utilized in confocal mode in order to limit the depth of analysis or to examine buried layers in a material, assuming that the visible light laser can penetrate to the buried layer in question. In confocal microscopy, the Raman scattering signal originating from above and below the focal plane will not reach the detector due to a pinhole aperture in front of the detector that blocks these out-of-focus light rays. Only the Raman signal from within the plane of focus (~2 μm depth resolution) will reach the detector [313].

Strengths and Weaknesses. In general, Raman spectroscopy combines all of the advantages associated with FTIR for chemical analysis with a significantly better lateral resolution. The ability to focus the visible light laser to a spot size of approximately 1 μm makes the technique ideal for mapping strain,

or other differences manifested in vibrational spectra, over surfaces. For stress/strain measurements, the technique is non-destructive and does not require sample preparation.

However, not all molecular vibrations are Raman active. A vibration must cause a change in the polarizability of the molecule in order to produce a Raman signal. In addition, fluorescent materials are difficult or impossible to analyze because the weak Raman signal can be inundated by the laser-induced fluorescent background. Since a powerful laser source is necessary in order to detect the weak Raman scattering bands, localized heating and/or photodecomposition can occur over the course of an analysis. Raman has limited depth resolution, typically about 2 μm in confocal mode; however, for some strongly absorbing materials, the depth resolution can be smaller, in the submicron to nanometer range.

References

1. *Materials Handbook Nineth Edition: Volume 10 Materials Characterization*. Metals Park, OH: American Society for Metals, 1986.
2. *Encyclopedia of Materials Characterization*. Boston: Butterworth-Heinemann, 1992.
3. Schroder, D. K. *Semiconductor Material and Device Characterization*. 3rd ed., New York: Wiley, 2006
4. Runyan, W. R., and T. J. Shaffner. *Semiconductor Measurements and Instrumentation*. New York: McGraw-Hill, 1998.
5. *Secondary Ion Mass Spectrometry: SIMS IX*. New York: Wiley, 1994.
6. *Electron Microscopy Society of America*. EMSA, 1998.
7. *47th Annual Denver X-Ray Conference*. Newtown Square, PA: International Centre for Diffraction Data, 1998.
8. Diebold, A., K. Shih, R. Colton, and J. Dagata, eds. *Industrial Applications of Scanned Probe Microscopy*. Gaithersburg, MD: NIST, 1994.
9. Seiler, D. G. *International Conference on Characterization and Metrology for ULSI Technology*. Woodbury, NY: American Institute of Physics, 1998.
10. Worledge, D. C. "Reduction of Positional Errors in a Four-Point Probe Resistance Measurement." *Appl. Phys. Lett.* 84 (2004): 1695–7.
11. Ishikawa, M., M. Yoshimura, and K. Ueda. "Development of Four-Probe Microscopy for Electric Conductivity Measurement." *Jpn J. Appl. Phys.* 44 (2005): 1502–3.
12. Albers, J., and H. L. Berkowitz. "An Alternative Approach to the Calculation of Four-Probe Resistances on Nonuniform Structures." *J. Electrochem. Soc.* 132 (1985): 2453–6 Weller, R. A. "An Algorithm for Computing Linear Four-Point Probe Thickness Correction Factors." *Rev. Sci. Instrum.* 72 (2001): 3580–6.
13. Perloff, D. S., J. N. Gan, and F. E. Wahl. "Dose Accuracy and Doping Uniformity of Ion Implantation Equipment." *Solid State Technol.* 24 (1981): 112–20; "ASTM Standard F1529-94 Standard Method for Sheet Resistance Uniformity by In-Line Four-Point Probe With the Dual-Configuration Procedure." *1996 Annual Book of ASTM Standards*, West Conshohocken, PA: American Society for Testing and Materials, 1996.
14. van der Pauw, L. J. "A Method of Measuring Specific Resistivity and Hall Effect of Discs of Arbitrary Shape." *Philips Res. Rep.* 13 (1958): 1–9 van der Pauw, L. J. "A Method of Measuring the Resistivity and Hall Coefficient on Lamellae of Arbitrary Shape." *Philips Tech. Rev.* 20 (1958): 220–4.
15. Versnel, W. "Analysis of Symmetrical van der Pauw Structures with Finite Contacts." *Solid-State Electron.* 21 (1978): 1261–8 Chwang, R., B. J. Smith, and C. R. Crowell. "Contact Size Effects on the van der Pauw Method for Resistivity and Hall Coefficient Measurement." *Solid-State Electron.* 17 (1974): 1217–27.
16. Sun, Y., J. Shi, and Q. Meng. "Measurement of Sheet Resistance of Cross Microareas Using a Modified van der Pauw Method." *Semicond. Sci. Technol.* 11 (1996): 805–11.
17. Buehler, M. G., and W. R. Thurber. "An Experimental Study of Various Cross Sheet Resistor Test Structures." *J. Electrochem. Soc.* 125 (1978): 645–50.

18. Buehler, M. G., S. D. Grant, and W. R. Thurber. "Bridge and van der Pauw Sheet Resistors for Characterizing the Line Width of Conducting Layers." *J. Electrochem. Soc.* 125 (1978): 650–4.
19. Buehler, M. G., and C. W. Hershey. "The Split-Cross-Bridge Resistor for Measuring the Sheet Resistance, Linewidth, and Line Spacing of Conducting Layers." *IEEE Trans. Elect. Dev.* ED-33 (1986): 1572–9; "ASTM Standard F1261M-95 Standard Test Method for Determining the Average Electrical Width of a Straight, Thin-Film Metal Line." *1996 Annual Book of ASTM Standards*, West Conshohocken, PA: American Society for Testing and Materials, 1996.
20. Cresswell, M. W., J. J. Sniegowski, R. N. Goshtagore, R. A. Allen, W. F. Guthrie, and L. W. Linholm. "Electrical Linewidth Test Structures Fabricated in Mono-Crystalline Films for Reference-Material Applications." In *Proceedings of the International Conference on Microelectronic Test Structures*, 16–24. Monterey, CA, 1997.
21. Chang, R., Y. Cao, and C. J. Spanos. "Modeling the Electrical Effects of Metal Dishing due to CMP for On-Chip Interconnect Optimization." *IEEE Trans. Elect. Dev.* 51 (2004): 1577–83.
22. Allen, R. A., M. W. Cresswell, and L. M. Buck. "A New Test Structure for the Electrical Measurement of the Width of Short Features with Arbitrarily Wide Voltage Taps." *IEEE Elect. Dev. Lett.* 13 (1992): 322–4.
23. Storms, G., S. Cheng, and I. Pollentier. "Electrical Linewidth Metrology for Sub-65 nm Applications." *Proc. SPIE* 5375 (2004): 614–28.
24. Rosencwaig, A. "Thermal-Wave Imaging." *Science* 218 (1982): 223–8.
25. Smith, W. L., A. Rosencwaig, and D. L. Willenborg. "Ion Implant Monitoring with Thermal Wave Technology." *Appl. Phys. Lett.* 47 (1985): 584–6; Smith, W. L., A. Rosencwaig, and D. L. Willenborg. "Ion Implant Monitoring with Thermal Wave Technology." *Solid-State Technol.* 29 (1986): 85–92.
26. Schroder, D. K. *Semiconductor Material and Device Characterization*. 3rd ed., New York: Wiley, 2006.
27. van Gelder, W., and E. H. Nicollian. "Silicon Impurity Distribution as Revealed by Pulsed MOS *C–V* Measurements." *J. Electrochem. Soc.* 118 (1971): 138–41.
28. Johnson, W. C., and P. T. Panousis. "The Influence of Debye Length on the *C–V* Measurement of Doping Profiles." *IEEE Trans. Elect. Dev.* ED-18 (1971): 965–73.
29. Barna, G. G., B. Van Eck, and J. W. Hosch. "In Situ Metrology." In *Handbook of Silicon Semiconductor Technology*, edited by A. C. Diebold, New York: Dekker, 2001.
30. Rommel, M. Semitest Inc., private correspondence.
31. Woolford, K., L. Newfield, and C. Panczyk. Monitoring Epitaxial Resistivity Profiles without Wafer Damage. Micro, July/August, (www.micromagazine.com) 2002.
32. Huang, Y., and C. C. Williams. "Capacitance–Voltage Measurement and Modeling on a Nanometer Scale by Scanning C–V Microscopy." *J. Vac. Sci. Technol.* B12 (1994): 369–72.
33. Neubauer, G., A. Erickson, C. C. Williams, J. J. Kopanski, M. Rodgers, and D. Adderton. "Two-Dimensional Scanning Capacitance Microscopy Measurements of Cross-Sectioned Very Large Scale Integration Test Structures." *J. Vac. Sci. Technol.* B14 (1996): 426–32 McMurray, J. S., J. Kim, and C. C. Williams. "Quantitative Measurement of Two-Dimensional Dopant Profile by Cross-sectional Scanning Capacitance Microscopy." *J. Vac. Sci. Technol.* B15 (1997): 1011–4.
34. Nakakura, C. Y., P. Tangyunyong, D. L. Hetherington, and M. R. Shaneyfelt. "Method for the Study of Semiconductor Device Operation Using Scanning Capacitance Microscopy." *Rev. Sci. Instrum.* 74 (2003): 127–33.
35. Williams, C. C. "Two-Dimensional Dopant Profiling by Scanning Capacitance Microscopy." *Annu. Rev. Mater. Sci.* 29 (1999): 471–504.
36. Vandervorst, W., P. Eyben, S. Callewaert, T. Hantschel, N. Duhayon, M. Xu, T. Trenkler, and T. Clarysse. "Towards Routine, Quantitative Two-Dimensional Carrier Profiling with Scanning Spreading Resistance Microscopy." In *Characterization and Metrology for ULSI Technology*, edited by D. G. Seiler, A. C. Diebold, T. J. Shaffner, R. McDonald, W. M. Bullis, P. J. Smith, and E. M. Secula, *American Institute of Physics*, Vol. 550, 613–9, 2000.
37. Eyben, P., N. Duhayon, D. Alvarez, and W. Vandervorst. "Assessing the Resolution Limits of Scanning Spreading Resistance Microscopy and Scanning Capacitance Microscopy." In

Characterization and Metrology for VLSI Technology: 2003 International Conference, edited by D. G. Seiler,
A. C. Diebold, T. J. Shaffner, R. McDonald, S. Zollner, R. P. Khosla, and E. M. Secula, *American Institute of Physics*, Vol. 683, 678–84, 2003.

38. Eyben, P., S. Denis, T. Clarysse, and W. Vandervorst. "Progress towards a Physical Contact Model for Scanning Spreading Resistance Microscopy." *Mater. Sci. Eng.* B102 (2003): 132–7.
39. "ASTM Standard F617M-95: Standard Method for Measuring MOSFET Linear Threshold Voltage." In *1996 Annual Book of ASTM Standards*, Conshohocken, PA: American Society for Testing and Materials, 1996.
40. Wong, H. S., M. H. White, T. J. Krutsick, and R. V. Booth. "Modeling of Transconductance Degradation and Extraction of Threshold Voltage in Thin Oxide MOSFETs." *Solid-State Electron.* 30 (1987): 953–68.
41. Jain, S. "Measurement of Threshold Voltage and Channel Length of Submicron MOSFETs." *Proc. IEE* 135, no. Pt I (1988): 162–4.
42. Ghibaudo, G. "New Method for the Extraction of MOSFET Parameters." *Electron. Lett.* 24 (1988): 543–5.
43. Ortiz-Conde, A., F. J. Garcia Sanchez, J. J. Liou, A. Cerdeira, M. Estrada, and Y. Yue. "A Review of Recent MOSFET Threshold Voltage Extraction Methods." *Microelectron. Reliab.* 42 (2002): 583–96.
44. Terada, K., and K.-I. Nishiyama. "Comparison of MOSFET-Threshold-Voltage Extraction Methods." *Solid-State Electron.* 45 (2001): 35–40.
45. Klaassen, F. M., and W. Hes. "On the Temperature Coefficient of the MOSFET Threshold Voltage." *Solid-State Electron.* 29 (1986): 787–9.
46. Ng, K. K., and J. R. Brews. "Measuring the Effective Channel Length of MOSFETs." *IEEE Circ. Dev.* 6 (1990): 33–8; McAndrew, C. C., and P. A. Layman. "MOSFET Effective Channel Length, Threshold Voltage, and Series Resistance Determination by Robust Optimization." *IEEE Trans. Elect. Dev.* 39 (1992): 2298–311.
47. Terada, K., and H. Muta. "A New Method to Determine Effective MOSFET Channel Length." *Jpn J. Appl. Phys.* 18 (1979): 953–9; Chern, J. G. J., P. Chang, R. F. Motta, and N. Godinho. "A New Method to Determine MOSFET Channel Length." *IEEE Elect. Dev. Lett.* EDL-1 (1980): 170–3.
48. Laux, S. E. "Accuracy of an Effective Channel Length/External Resistance Extraction Algorithm for MOSFET's." *IEEE Trans. Elect. Dev.* ED-31 (1984): 1245–51.
49. De La Moneda, F. H., H. N. Kotecha, and M. Shatzkes. "Measurement of MOSFET Constants." *IEEE Elect. Dev. Lett.* EDL-3 (1982): 10–2.
50. Taur, Y., D. S. Zicherman, D. R. Lombardi, P. R. Restle, C. H. Hsu, H. I. Hanafi, M. R. Wordeman, B. Davari, and G. G. Shahidi. "A New "Shift and Ratio" Method for MOSFET Channel-Length Extraction." *IEEE Elect. Dev. Lett.* 13 (1992): 267–9.
51. Takeda, E., C. Y. Yang, and A. Miura-Hamada. *Hot Carrier Effects in MOS Devices*. San Diego, CA: Academic Press, 1995 Acovic, A., G. La Rosa, and Y. C. Sun. "A Review of Hot-Carrier Degradation Mechanisms in MOSFETs." *Microelectron. Reliab.* 36 (1996): 845–69.
52. Chang, W. H., B. Davari, M. R. Wordeman, Y. Taur, C. C. H. Hsu, and M. D. Rodriguez. "A High-Performance 0.25 μm CMOS Technology." *IEEE Trans. Elect. Dev.* 39 (1992): 959–66.
53. Yue, J. T. "Reliability." In *ULSI Technology*, edited by C. Y. Chang, and S. M. Sze, New York: McGraw-Hill, 1996.
54. Li, E., E. Rosenbaum, J. Tao, and P. Fang. "Projecting Lifetime of Deep Submicron MOSFETs." *IEEE Trans. Elect. Dev.* 48 (2001): 671–8; Cheng, K., and J. W. Lyding. "An Analytical Model to Project MOS Transistor Lifetime Improvement by Deuterium Passivation of Interface Traps." *IEEE Elect. Dev. Lett.* 24 (2003): 655–7.
55. Shin, H. C., and C. M. Hu. "Dependence of Plasma-Induced Oxide Charging Current on Al Antenna Geometry." *IEEE Elect. Dev. Lett.* 13 (1992): 600–2; Eriguchi, K., Y. Uraoka, H. Nakagawa, T. Tamaki, M. Kubota, and N. Nomura. "Quantitative Evaluation of Gate Oxide Damage during Plasma Processing Using Antenna Structure Capacitors." *Jpn J. Appl. Phys.* 33 (1994): 83–7.

56. Shideler, J., S. Reno, R. Bammi, C. Messick, A. Cowley, and W. Lukas. "A New Technique for Solving Wafer Charging Problems." *Semicond. Int.* 18 (1995): 153–8; Lukaszek, W. "Understanding and Controlling Wafer Charging Damage." *Solid State Technol.* 41 (1998): 101–12.
57. Kuhn, M., and D. J. Silversmith. "Ionic Contamination and Transport of Mobile Ions in MOS Structures." *J. Electrochem. Soc.* 118 (1971): 966–70; Hillen, M. W., and J. F. Verwey. "Mobile Ions in SiO_2 Layers on Si." In *Instabilities in Silicon Devices: Silicon Passivation and Related Instabilities*, edited by G. Barbottin, and A. Vapaille, 403–39. Amsterdam: Elsevier, 1971.
58. Stauffer, L., T. Wiley, T. Tiwald, R. Hance, P. Rai-Choudhury, and D. K. Schroder. "Mobile Ion Monitoring by Triangular Voltage Sweep." *Solid-State Technol.* 38 (1995): S3–S8.
59. Nicollian, E. H., and J. R. Brews. *MOS Physics and Technology*. New York: Wiley, 1982.
60. Ng, K. K. *Complete Guide to Semiconductor Devices*. 2nd ed, 183. Wiley-InterScience: New York, 2002.
61. Saks, N. S., and M. G. Ancona. "Determination of Interface Trap Capture Cross Sections Using Three-Level Charge Pumping." *IEEE Elect. Dev. Lett.* 11 (1990): 339–41; Siergiej, R. R., M. H. White, and N. S. Saks. "Theory and Measurement of Quantization Effects on Si–SiO_2 Interface Trap Modeling." *Solid-State Electron.* 35 (1992): 843–54.
62. Yoneda, K., K. Okuma, K. Hagiwara, and Y. Todokoro. "The Reliability Evaluation of Thin Silicon Dioxide Using the Stepped Current TDDB Technique." *J. Electrochem. Soc.* 142 (1995): 596–600.
63. Wolters, D. R., and J. R. Verwey. "Breakdown and Wear-Out Phenomena in SiO_2 Films." In *Instabilities in Silicon Devices*, edited by B. Barbottin, and A. Vapaille, 315–62. Amsterdam: North-Holland, 1986; Wolters, D. R. "Breakdown and Wearout Phenomena in SiO_2." In *Insulating Films on Semiconductors*, edited by M. Schulz, and G. Pensl, 180–94. Berlin: Springer, 1986.
64. Wolters, D. R., and J. J. van der Schoot. "Dielectric Breakdown in MOS Devices." *Philips J. Res.* 40 (1985): 115–92.
65. Lang D. V. "Deep-Level Transient Spectroscopy: A New Method to Characterize Traps in Semiconductors." *J. Appl. Phys.* 45 (1974): 3023–32; Lang, D. V. "Fast Capacitance Transient Apparatus: Application to ZnO and O Centers in GaP p–n Junctions." *J. Appl. Phys.* 45 (1974): 3014–22. ASTM Standard F 978-90; Miller, G. L., D. V. Lang, and L. C.Kimerling. "Capacitance Transient Spectroscopy." In *Annual Review Material Science*, Vol. 7, edited by R. A. Huggins, R. H. Bube, and R. W. Roberts, *Annual Review Material Science*, 377–448. Palo Alto, CA: Annual Reviews, 1974.
66. M'saad, H., J. Michel, J. J. Lappe, and L. C. Kimerling. "Electronic Passivation of Silicon Surfaces by Halogens." *J. Electron. Mat.* 23 (1994): 487–91.
67. Zoth, G., and W. Bergholz. "A Fast, Preparation-Free Method to Detect Iron in Silicon." *J. Appl. Phys.* 67 (1990): 6764–71.
68. Schroder, D. K. "The Concept of Generation and Recombination Lifetimes in Semiconductors." *IEEE Trans. Elect. Dev.* ED-29 (1982): 1336–8.
69. Obermeier, G., and D. Huber. "Iron Detection in Polished and Epitaxial Wafers Using Generation Lifetime Measurements." *J. Appl. Phys.* 81 (1997): 7345–9.
70. Lee, S. Y., and D. K. Schroder. "Measurement Time Reduction for Generation Lifetimes." *IEEE Trans. Elect. Dev.* 46 (1999): 1016–21.
71. Schroder, D. K. *Semiconductor Material and Device Characterization*. 3rd ed., New York: Wiley, 2006.
72. Schroder, D. K., M. S. Fung, R. L. Verkuil, S. Pandey, W. H. Howland, and M. Kleefstra. "Corona-Oxide-Semiconductor Generation Lifetime Characterization." *Solid-State Electron.* 42 (1998): 505–12.
73. Grove, A. S., and D. J. Fitzgerald. "Surface Effects on pn Junctions: Characteristics of Surface Space-Charge Regions Under Non-Equilibrium Conditions." *Solid-State Electron.* 9 (1966): 783–806; Fitzgerald, D. J., and A. S. Grove. "Surface Recombination in Semiconductors." *Surf. Sci.* 9 (1968): 347–69.
74. Fung, M. S., and R. L. Verkuil. "Contactless Measurement of Silicon Generation Leakage and Crystal Defects by a Corona-Pulsed Deep-Depletion Potential Transient Technique." *Extended Abstracts*, Chicago, IL; The Electrochemical Society Meet, 1988; Verkuil, R. L., and M. S. Fung. "Contactless Silicon Doping Measurements by Means of a Corona-Oxide-Semiconductor (COS)

Technique." *Extended Abstracts*, Chicago, IL: The Electrochemical Society Meet, 1988; Fung, M. S., and R. L. Verkuil. "Process Learning by Nondestructive Lifetime Testing." In *Semiconductor Silicon 1990*, edited by H. R. Huff, K. G. Barraclough, and J. I. Chikawa, 924–50. Pennington, NJ: The Electrochemical Society, 1990; Verkuil, R. L., and M. S. Fung. "A Contactless Alternative to MOS Charge Measurements by Means of a Corona-Oxide-Semiconductor (COS) Technique." *Extended Abstracts*, Chicago, IL: The Electrochemical Society Meet, 1988.

75. Kelvin, L. "On a Method of Measuring Contact Electricity." *Nature* (1881) Kelvin, L. "Contact Electricity of Metals." *Philos. Mag.* 46 (1898): 82–121.
76. Kronik, L., and Y. Shapira. "Surface Photovoltage Phenomena: Theory, Experiment, and Applications." *Surf. Sci. Rep.* 37 (1999): 1–206.
77. Williams, R., and M. H. Woods. "High Electric Fields in Silicon Dioxide Produced by Corona Charging." *J. Appl. Phys.* 44 (1973): 1026–8 Weinberg, Z. A. "Tunneling of Electrons from Si into Thermally Grown SiO_2." *Solid-State Electron.* 20 (1977): 11–18; Woods, M. H., and R. Williams. "Injection and Removal of Ionic Charge at Room Temperature through the Interface of Air with SiO_2." *J. Appl. Phys.* 44 (1973): 5506–10.
78. Schroder, D. K. "Surface Voltage and Surface Photovoltage: History, Theory and Applications." *Meas. Sci. Technol.* 12 (2001): R16–R31; Schroder, D. K. "Contactless Surface Charge Semiconductor Characterization." *Mater. Sci. Eng.* B91–92 (2002): 196–210.
79. Weinzierl, S. R., and T. G. Miller. "Non-Contact Corona-Based Process Control Measurements: Where We've Been and Where We're Headed." In *Analytical and Diagnostic Techniques for Semiconductor Materials, Devices, and Processes*, edited by B. O. Kolbesen, C. Claeys, P. Stallhofer, F. Tardif, J. Benton, T. Shaffner, D. Schroder, P. Kishino, and P. Rai-Choudhury, 342–50. Pennington, NJ: The Electrochemical Society, 1999 (ECS 99-16).
80. Roy, P. K., C. Chacon, Y. Ma, I. C. Kizilyalli, G. S. Horner, R. L. Verkuil, and T. G. Miller. "Non-Contact Characterization of Ultrathin Dielectrics for the Gigabit Era." In *Diagnostic Techniques for Semiconductor Materials and Devices*, edited by P. Rai-Choudhury, J. L. Benton, D. K. Schroder, and T. J. Shaffner, 280–94. Pennington, NJ: The Electrochemical Society, 1997 (PV97-12).
81. Miller, T. G. "A New Approach for Measuring Oxide Thickness." *Semicond. Int.* 18 (1995): 147–8.
82. Lo, S. H., D. A. Buchanan, and Y. Taur. "Modeling and Characterization of Quantization, Polysilicon Depletion, and Direct Tunneling Effects in MOSFETs With Ultrathin Oxides." *IBM J. Res. Dev.* 43 (1999): 327–37.
83. Weinberg, Z. A., W. C. Johnson, and M. A. Lampert. "High-Field Transport in SiO_2 on Silicon Induced by Corona Charging of the Unmetallized Surface." *J. Appl. Phys.* 47 (1976): 248–55.
84. Bonnell, D. A. *Scanning Probe Microscopy and Spectroscopy*. 2nd ed. New York: Wiley-VCH, 2001.
85. Shaffner, T. J. "Characterization Challenges for the ULSI Era." In *Diagnostic Techniques for Semiconductor Materials and Devices*, edited by P. Rai-Choudhury, J. L. Benton, D. K. Schroder, and T. J. Shaffner, 1–15. Pennington, NJ: The Electrochemical Society, 1997.
86. Hamers, R. J., and D. F. Padowitz. "Methods of Tunneling Spectroscopy with the STM." In *Scanning Probe Microscopy and Spectroscopy*, 2nd ed., edited by D. Bonnell, New York: Wiley-VCH, 2001 (chap. 4).
87. Smith, R. L., and G. S. Rohrer. "The Preparation of Tip and Sample Surfaces for Scanning Probe Experiments." In *Scanning Probe Microscopy and Spectroscopy*, 2nd ed., edited by D. Bonnell, New York: Wiley-VCH, 2001 (chap. 6).
88. Meyer, E., H. J. Hug, and R. Bennewitz. *Scanning Probe Microscopy*. Berlin: Springer, 2004.
89. Simmons, J. "Generalized Formula for the Electric Tunnel Effect between Similar Electrodes Separated by a Thin Insulating Film." *J. Appl. Phys.* 34 (1963): 1793–803.
90. Binnig, G., C. F. Quate, and C. H. Gerber. "Atomic Force Microscope." *Phys. Rev. Lett.* 56 (1986): 930–3.
91. Quate, C. F. "The AFM as a Tool for Surface Imaging." *Surf. Sci.* 299–300 (1994): 980–95.
92. Sarid, D. *Scanning Force Microscopy with Applications to Electric, Magnetic, and Atomic Forces*. Revised Edition New York: Oxford University Press, 1994.
93. Meyer, G., and N. M. Amer. "Novel Optical Approach to Atomic Force Microscopy." *Appl. Phys. Lett.* 53 (1988): 045–1047.

94. Zhong, Q., D. Inniss, K. Kjoller, and V. B. Elings. "Fractured Polymer/Silica Fiber Surface Studied by Tapping Mode Atomic Force Microscopy." *Surf. Sci. Lett.* 290 (1993): L668–L92.
95. Nonnenmacher, M., M. P. Boyle, and H. K. Wickramasinghe. "Kelvin Probe Microscopy." *Appl. Phys. Lett.* 58 (1991): 2921–3.
96. Weaver, J. M. R., and H. K. Wickramasinghe. "Semiconductor Characterization by Scanning Force Microscope Surface Photovoltage Microscopy." *J. Vac. Sci. Technol.* B9 (1991): 1562–5.
97. Bonnell, D. A., and S. Kalinin. "Local Potential at Atomically Abrupt Oxide Grain Boundaries by Scanning Probe Microscopy." In *Proceedings of the International Meeting on Polycrystalline Semiconductors*, edited by O. Bonnaud, T. Mohammed-Brahim, H. P. Strunk, and J. H. Werner, *Solid State Phenomena*, 33–47. Switzerland: Scitech Publ. Uettikon am See, 2001.
98. Goldstein, J. I., D. E. Newbury, P. Echlin, D. C. Joy, C. Fiori, and E. Lifshin. *Scanning Electron Microscopy and X-Ray Microanalysis*. New York: Plenum Press, 1981.
99. Verhoeven, J. D. "Scanning Electron Microscopy." In *Metals Handbook, Ninth Edition: Volume 10 Materials Characterization*, edited by R. E. Whan, K. Mills, J. R. Davis, J. D. Destefani, D. A. Dieterich, G. M. Crankovic, H. J. Frissell, D. M. Jenkins, W. H. Cubberly, and R. L. Stedfeld, 490–515. Metals Park, OH: American Society for Metals, 1986.
100. Wells, O. C., A. Boyde, E. Lifshin, and A. Rezanowich. *Scanning Electron Microscopy*. New York: McGraw-Hill, 1974.
101. Murr, L. E. *Scanning Electron Microscopy*. New York: Marcel Dekker, Inc., 1982.
102. Goldstein, J. I., and H. Yakowitz. *Practical Scanning Electron Microscopy*. New York: Plenum Press, 1975.
103. Joy, D. C., A. D. Romig Jr.., and J. I. Goldstein. *Principles of Analytical Electron Microscopy*. New York: Plenum Press, 1986.
104. Belcher, R. W., G. P. Hart, and W. R. Wade. "Preparation of Semiconductor Devices for Scanning Electron Microscopy and Quantification Using Etchback Methods." *Scanning Electron Microsc.* II (1984): 613–24.
105. Koellen, D. S., D. I. Saxon, and K. E. Wendel. "Cross-Sectional Analysis of Silicon Metal Oxide Semiconductor Devices Using the Scanning Electron Microscope." *Scanning Electron Microsc.* I (1985): 43–53.
106. Mills, T. "Precision VLSI Cross-Sectioning and Staining." *Proc. IEEE: Rel. Phys.* (1983): 324–31.
107. Angelides, P. G. "Precision Cross Sectional Analysis of LSI and VLSI Devices." *Proc. IEEE: Rel. Phys.* (1981): 134–8.
108. Hammond, B. R., and T. R. Vogel. "Non-Encapsulated Microsectioning as a Construction and Failure Analysis Technique." *Proc. IEEE: Rel. Phys.* (1982): 221–3.
109. Gill, M., and E. Woster. "The In-Fab SEM/EDX Integration Challenge." *Semicond. Int.* 16 (1993): 78–82.
110. Shaffner, T. J., and J. W. S. Hearle. "Recent Advances in Understanding Specimen Charging." *Scanning Electron Microsc.* I (1976): 61–82.
111. Kuroda, K., S. Hosoki, and T. Komoda. "Observation of Tungsten Field Emitter Tips with an Ultra-High Resoltuion Field Emission Scanning Electron Microscope." *Scanning Microsc.* 1 (1987): 911–7.
112. Kanaya, K., and S. Okayama. "Penetration and Energy-Loss Theory of Electrons in Solid Targets." *J. Phys. D: Appl. Phys.* 5 (1972): 43–58.
113. Everhart, T. E., and R. F. M. Thornley. "Wide-Band Detector for Micro-Microampere Low-Energy Electron Currents." *J. Sci. Instrum.* 37 (1960): 246–8.
114. McKinley, T. D., K. F. J. Heinrich, and D. B. Wittry. *The Electron Microprobe*. New York: Wiley, 1966.
115. Castaing, R. "Electron Probe Microanalysis." In *Advances in Electronics and Electron Physics*, edited by L. Marton, and C. Marton, 317–86. New York: Academic Press, 1960.
116. Reed, S. J. B. *Electron Microprobe Analysis*. New York: Cambridge University Press, 1993.
117. Fitzgerald, R., K. Keil, and K. F. J. Heinrich. "Solid-State Energy-Dispersion Spectrometer for Electron-Microprobe X-Ray Analysis." *Science* 159 (1968): 528–30.

118. Titchmarsh, J. M. "Energy Dispersive X-Ray Analysis (EDX) in the TEM/STEM." In *Quantitative Microbeam Analysis*, edited by A. G. Fitzgerald, B. E. Storey, and D. Fabian, 275–301. Bristol, UK: Institute of Physics, 1995.
119. Powell, C. J., and M. P. Seah. "Precision, Accuracy, and Uncertainty in Quantitative Surface Analyses by Auger-Electron Spectroscopy and X-Ray Photoelectron Spectroscopy." *J. Vac. Sci. Technol. A* 8 (1990): 735–63.
120. Seah, M. P., and G. C. Smith. "Quantitative AES and XPS: Calibration of Electron Spectrometers for True Spectral Measurements—VAMAS Round Robins and Parameters for Reference Spectral Data Banks." *Vacuum* 41 (1990): 1601–4.
121. Mroczkowski, S., and D. Lichtman. "Calculated Auger Yields and Sensitivity Factors for KLL-NOO Transitions with 1–10 kV Primary Beams." *J. Vac. Sci. Technol. A* 3 (1985): 1860–5.
122. Shaffner, T. J. "Surface Characterization for VLSI." In *Materials and Process Characterization*, edited by N. G. Einspruch, and G. B. Larrabee, 497–527. New York: Academic Press, 1983.
123. Aton, T. J., K. A. Joyner, C. H. Blanton, A. T. Appel, M. G. Harward, M. H. Bennett-Lilley, and S. S. Mahant-Shetti. "Using Scanning Electron Beams for Testing Microstructure Isolation and Continuity." *Proc. IEEE: Rel. Phys.* (1991): 239–44.
124. Menzel, E., and R. Buchanan. "Electron Beam Probing of Integrated Circuits." *Solid-State Technol.* 28 (1985): 63–70.
125. Todokoro, H., and S. Yoneda. "Electron Beam Tester with 10 ps Time Resolution." *Proc. Int. Test Conf.* (1986): 600–6.
126. Gonzales, A. J. "On the Electron Beam Induced Current Analysis of Semiconductor Devices." *Scanning Electron Microsc.* IV (1974): 941–8.
127. Bresse, J. F. "Electron Beam Induced Current in Silicon Planar p–n Junctions: Physical Model of Carrier Generation, Determination of Some Physical Parameters in Silicon." *Scanning Electron Microsc.* I (1972): 105–12.
128. Holt, D. B., and F. M. Saba. "The Cathodoluminescence Mode of the Scanning Electron Microscope: A Powerful Microcharacterization Technique." *Scanning Electron Microsc.* III (1985): 1023–45.
129. Yacobi, B. G., and D. B. Holt. "Cathodoluminescence Scanning Electron Microscopy of Semiconductors." *J. Appl. Phys.* 59 (1986): R1–R24.
130. Roedel, R. J., S. Myhajlenko, J. L. Edwards, and K. Rowley. "Cathodoluminescence Characterization of Semiconductor Materials." *Proc. Electrochem. Soc.* 88–20 (1988): 185–96.
131. Pfefferkorn, G., W. Brocker, and M. Hastenrath. "The Cathodoluminescence Method in the Scanning Electron Microscope." *Scanning Electron Microsc.* I (1980): 251–8.
132. Holt, D. B., and S. Datta. "The Cathodoluminescent Mode as an Analytical Technique: Its Development and Prospects." *Scanning Electron Microsc.* I (1980): 259–78.
133. Heard, P. "Cathodoluminescence—Interesting Phenomenon or Useful Technique?" *Microsc. Anal.* (1996): 25–7.
134. Link Analytical, AN10000 X-Ray Microanalysis System Bucks, U.K. 1994.
135. Dingley, D. J., and K. Baba-Kishi. "Use of Electron Back Scatter Diffraction Patterns for Determination of Crystal Symmetry Elements." *Scanning Electron Microsc.* II (1986): 383–91.
136. Moore, T. M., S. Matteson, W. M. Duncan, and R. J. Matyi. "Microstructural Characterization of GaAs Substrates." *Mat. Res. Soc. Symp. Proc.* 69 (1986): 379–84.
137. Amptek, I. X-Ray detector: XR-100T Bedford, MA. 1995.
138. Silver, E., M. LeGros, N. Madden, J. Beeman, and E. Haller. "High-Resolution, Broad-Band Microcalorimeters for X-Ray Microanalysis." *X-Ray Spectrom.* 25 (1996): 115–22.
139. Wollman, D. A., K. D. Irwin, G. C. Hilton, L. L. Dulcie, D. E. Newbury, and J. M. Martinis. "High-Resolution, Energy-Dispersive Microcalorimeter Spectrometer for X-Ray Microanalysis." *J. Microsc.* 188 (1997): 196–223.
140. Silver, E., and M. LeGros. "The Application of a High Resolution, Broad Band Microcalorimeter to the SEM-Based Microanalysis Problem." *X-Ray Spectrom.* (1995): 1–13.
141. Robinson, M. "A Microcalorimeter for High Resolution, Broad Band X-Ray Microanalysis 94-01." 1–24. Livermore, CA: Lawrence Livermore National Laboratory, 1994.

142. McCammon, D., W. Cui, M. Juda, J. Morgenthaler, J. Zhang, R. L. Kelley, S. S. Holt, G. M. Madejski, S. H. Moseley, and A. E. Szymkowiak. "Thermal Calorimeters for High Resolution X-Ray Spectroscopy." *Nucl. Instrum. Methods A* 326 (1993): 157–65.
143. LeGros, M., E. Silver, D. Schneider, J. McDonald, S. Bardin, R. Schuch, N. Madden, and J. Beeman. "The First High Resolution, Broad Band X-Ray Spectroscopy of Ion-Surface Interactions Using a Microcalorimeter." *Nucl. Instrum. Methods A* 357 (1996): 110–4.
144. Lesyna, L., D. D. Marzio, S. Gottesman, and M. Kesselman. "Advanced X-Ray Detectors for the Analysis of Materials." *J. Low Temp. Phys.* 93 (1993): 779–84.
145. Williams, D. B., and C. B. Carter. *Transmission Electron Microscopy*. New York: Plenum Press, 1996.
146. Reimer, L. *Transmission Electron Microscopy*. New York: Springer, 1989.
147. Williams, D. B. *Practical Analytical Electron Microscopy in Materials Science*. Mahwah, NJ: Philips Electronic Instruments, Inc., 1984.
148. Marcus, R. B., and T. T. Sheng. *Transmission Electron Microscopy of Silicon VLSI Circuits and Structures*. New York: Wiley, 1983.
149. Murr, L. E. *Electron and Ion Microscopy and Microanalysis*. New York: Marcel Dekker, Inc., 1982.
150. Buseck, P., J. Cowley, and L. Eyring, eds. *High-Resolution Transmission Electron Microscopy and Associated Techniques*. Oxford, U.K.: Oxford University Press, 1988.
151. Cullis, A. G., ed. *Microscopy of Semiconducting Materials 1991*. Bristol, U.K.: IOP Publishing, Ltd, 1991.
152. *Materials Problem Solving with the Transmission Electron Microscope*. Pittsburgh, PA: Materials Research Society, 1986.
153. *High-Resolution Transmission Electron Microscopy*. Oxford, U.K.: Oxford University Press, 1988.
154. Egerton, R. F. *Electron Energy Loss Spectroscopy in the Electron Microscope*. New York: Plenum Press, 1986.
155. Hall, C. E. *Introduction to Electron Microscopy*. New York: McGraw Hill, 1966.
156. *Specimen Preparation for Transmission Electron Microscopy of Materials IV*. Warrendale, PA: Materials Research Society, 1997.
157. *Specimen Preparation for Transmission Electron Microscopy of Materials*. Pittsburgh, PA: Materials Research Society, 1988.
158. *Specimen Preparation for Transmission Electron Microscopy of Materials II*. Pittsburgh, PA: Materials Research Society, 1990.
159. Giannuzzi, L. A., J. L. Drown, S. R. Brown, R. B. Irwin, and F. A. Stevie. "Focused Ion Beam Milling and Micromanipulation Lift-Out for Site Specific Cross-Section TEM Specimen Preparation." In *Specimen Preparation for Transmission Electron Microscopy of Materials IV*, edited by R. M. Anderson, and S. D. Walck, 19–27. Warrendale, PA: Materials Research Society, 1997.
160. Su, DH-I., H. T. Shishido, F. Tsai, L. Liang, and F. C. Mercado. "A Detailed Procedure for Reliable Preparation of TEM Samples Using FIB Milling." In *Specimen Preparation for Transmission Electron Microscopy of Materials IV*, edited by R. M. Anderson, and S. D. Walck, 105–16. Warrendale, PA: Materials Research Society, 1997.
161. Shaapur, F., T. Stark, T. Woodward, and R. J. Graham. "Evaluation of a New Strategy for Transverse TEM Specimen Preparation by Focused-Ion-Beam Thinning." In *Specimen Preparation for Transmission Electron Microscopy of Materials IV*, edited by R. M. Anderson, and S. D. Walck, 173–80. Warrendale, PA: Materials Research Society, 1997.
162. Tsujimoto, K., S. Tsuji, H. Takatsuji, K. Kuroda, H. Saka, and N. Miura. "Cross-Sectional TEM Sample Preparation Method Using FIB Etching for Thin-Film Transistor." In *Specimen Preparation for Transmission Electron Microscopy of Materials IV*, edited by R. M. Anderson, and S. D. Walck, 207–15. Warrendale, PA: Materials Research Society, 1997.
163. Pramanik, D., and J. Glanville. "Aluminum Film Analysis with the Focused Ion Beam Microscope." *Solid-State Technol.* 55 (1990): 77–80.
164. Gamo, K., N. Takakura, N. Samoto, R. Shimizu, and S. Namba. "Ion Beam Assisted Deposition of Metal Organic Films Using Focused Ion Beams." *Jpn J. Appl. Phys.* 23 (1984): L293–L5.
165. Shedd, G. M., H. Lezec, A. D. Dubner, and J. Melngailis. "Focused Ion Beam Induced Deposition of Gold." *Appl. Phys. Lett.* 49 (1986): 1584–6.

166. Mashiko, Y., H. Morimoto, H. Koyama, S. Kawazu, T. Kaito, and T. Adachi. "A New VLSI Diagnosis Technique: Focused Ion Beam Assisted Multi-Level Circuit Probing." *Proc. IEEE: Rel. Phys.* (1987): 111–7.
167. Matusiewicz, G. R., S. J. Kirch, V. J. Seeley, and P. G. Blauner. "The Role of Focused Ion Beams in Physical Failure Analysis." *Proc. IEEE: Rel. Phys.* (1991): 167–70.
168. Komano, H., Y. Ogawa, and T. Takigawa. "Silicon Oxide Film Formation by Focused Ion Beam (FIB)-Assisted Deposition." *Jpn J. Appl. Phys.* 28 (1989): 2372–5.
169. Binnig, G., and H. Rohrer. "The Scanning Tunneling Microscope." *Sci. Am.* 253 (1985): 50–6.
170. Binnig, G., and H. Rohrer. "Scanning Tunneling Microscopy—From Birth to Adolescence." *Rev. Mod. Phys.* 59 (1987): 615–25.
171. Binnig, G., H. Rohrer, C. Gerber, and E. Weibel. "Surface Studies by Scanning Tunneling Microscopy." *Phys. Rev. Lett.* 49 (1982): 57–61.
172. Wickramasinghe, H. K. "Scanned-Probe Microscopes." *Sci. Am.* October (1989): 98–105.
173. Chen, C. J. *Introduction to Scanning Tunneling Microscopy.* Oxford, U.K.: Oxford University Press, 1993.
174. Marrian, C. R. K., ed. *Technology of Proximal Probe Lithography.* Bellingham, WA: SPIE Optical Engineering Press, 1993.
175. STM '90, The Fifth International Conference on Scanning Tunneling Microscopy/Spectroscopy and NANO I, The First International Conference on Nanometer Scale Science and Technology. American Vacuum Society, 1990.
176. Bell, L. D., and W. J. Kaiser. "Imaging Subsurface Interfaces by Ballistic-Electron-Emission Microscopy." In *Diagnostic Techniques for Semiconductor Materials and Devices*, edited by T. J. Shaffner, and D. K. Schroder, 97–108. Pennington, NJ: The Electrochemical Society, 1988.
177. Baum, R. "Chemical Force Microscopy: Method Maps Functional Groups on Surfaces." *Chem. Eng. News* 73 (1994): 6.
178. Houston, J. E., and T. A. Michalske. "The Interfacial-Force Microscope." *Nature* 356 (1992): 266–7.
179. Babcock, K. "Magnetic Force Microscopy." *Photon. Spectra* (1994).
180. Sidles, J. A., J. L. Garbini, K. J. Bruland, D. Rugar, O. Zuger, S. Hoen, and C. S. Yannoni. "Magnetic Resonance Force Microscopy." *Rev. Mod. Phys.* 67 (1995): 249–65.
181. Wago, K., O. Zuger, R. Kendrick, C. S. Yannoni, and D. Rugar. "Low-Temperature Magnetic Resonance Force Detection." *J. Vac. Sci. Technol. B* 14 (1996): 1197–201.
182. Electric field Measurements with the MultiMode™ AFM Santa Barbara, CA: Digital Instruments Inc., 1996.
183. Rugar, D., O. Zuger, S. Hoen, C. S. Yannoni, H.-M. Vieth, and R. D. Kendrick. "Force Detection of Nuclear Magnetic Resonance." *Science* 264 (1994): 1560–3.
184. Thompson, C. A., R. W. Cross, and A. B. Kos. "Micromagnetic Scanning Microprobe System." *Rev. Sci. Instrum.* 65 (1994): 383–9.
185. Vandervorst, W., T. Clarysse, P. De Wolf, L. Hellemans, J. Snauwaert, V. Privitera, and V. Raineri. "On the Determination of Two-Dimensional Carrier Distributions." *Nucl. Instrum. Methods B* 96 (1995): 123–32.
186. Clarysse, T., P. De Wolf, H. Bender, and W. Vandervorst. "Recent Insights into the Physical Modeling of the Spreading Resistance Point Contact." *J. Vac. Sci. Technol. B* 14 (1996): 358–68.
187. Shafai, C., D. J. Thomson, M. Simard-Normandin, G. Mattiussi, and P. J. Scanlon. "Delineation of Semiconductor Doping by Scanning Resistance Microscopy." *Appl. Phys. Lett.* 64 (1994): 342–4.
188. Paesler, M. A., and P. J. Moyer. *Near-Field Optics Theory, Instrumentation, and Applications*. New York: Wiley, 1996.
189. Betzig, E., and J. K. Trautman. "Near-Field Optics: Microscopy, Spectroscopy, and Surface Modification beyond the Diffraction Limit." *Science* 257 (1992): 189–96.
190. Duncan, W. M. "Near-Field Scanning Optical Microscope for Microelectronic Materials and Devices." *J. Vac. Sci. Technol. A* 14 (1996): 1914–8.
191. Kopanski, J. J., J. F. Marchiando, and J. R. Lowney. "Scanning Capacitance Microscopy Measurements and Modeling: Progress towards Dopant Profiling of Silicon." *J. Vac. Sci. Technol. B* 14 (1996): 242–7.

192. Erickson, A., L. Sadwick, G. Neubauer, J. Kopanski, D. Adderton, and M. Rogers. "Quantitative Scanning Capacitance Microscopy Analysis of Two-Dimensional Dopant Concentrations at Nanoscale Dimensions." *J. Electron. Mat.* 25 (1996): 301–4.
193. Neubauer, G., A. Erickson, C. C. Williams, J. J. Kopanski, M. Rodgers, and D. Adderton. "Two-Dimensional Scanning Capacitance Microscopy Measurements of Cross-Sectioned Very Large Scale Integration Test Structures." *J. Vac. Sci. Technol. B* 14 (1996): 426–32.
194. Krieger, J. "Use of Scanning Probe Microscopy Expanding." *Chem. Eng. News* (1993): 30–1.
195. Liu, H., F. F. Fan, W. Lin, and A. J. Bard. "Scanning Electrochemical and Tunneling Ultramicroelectrode Microscope for High-Resolution Examination of Electrode Surfaces in Solution." *J. Am. Chem. Soc.* 108 (1986): 3838–9.
196. Hochwitz, T., A. K. Henning, C. Levey, C. Daghlian, J. Slinkman, J. Never, P. Kaszuba, R. Gluck, R. Wells, and J. Pekarik. "Imaging Integrated Circuit Dopant Profiles with the Force-Based Scanning Kelvin Probe Microscope." *J. Vac. Sci. Technol. B* 14 (1996): 440–6.
197. Wessels, B. W., and L. Q. Qian. "Scanning Tunneling Optical Spectroscopy of Semiconductor Thin Films and Quantum Wells." *J. Vac. Sci. Technol. B* 10 (1992): 1803–6.
198. Shluger, A., C. Pisani, C. Roetti, and R. Orlando. "Ab Initio Simulation of the Interaction between Ionic Crystal Surfaces and the Atomic Force Microscope Tip." *J. Vac. Sci. Technol. A* 8 (1990): 3967–72.
199. Ivanov, G. K., M. A. Kozhushner, and I. I. Oleinik. "Direct and Inverse Problems in the Theory of Scanning Tunneling Microscopy." *Surf. Sci.* 331–333 (1995): 1191–6.
200. Chen, C. J. "In-Situ Characterization of Tip Electronic Structure in Scanning Tunneling Microscopy." *Ultramicroscopy* 42–44 (1992): 147–53.
201. Keller, D. "Reconstruction of STM and AFM Images Distorted by Finite-Size Tips." *Surf. Sci.* 253 (1991): 353–64.
202. Griffith, J. E., D. A. Grigg, G. P. Kochanski, M. J. Vasile, and P. E. Russell. "Metrology with Scanning Probe Microscopes." In *Technology of Proximal Probe Lithography*, edited by C. R. K. Marian, 364–89. Bellingham, WA: SPIE Optical Engineering Press, 1993.
203. Benninghoven, A., J. L. Hunter, Jr., B. W. Schueler, H. H. Smith, H. W. Werner, eds. "Secondary Ion Mass Spectrometry: SIMS XIV, (San Diego)." *Applied Surface Science*, Vol. 231–232, 475–478. North-Holland: Elsevier, 2004.
204. Benninghoven, A., Y. Nihei, M. Kudo, Y. Homma, H. Yurimoto, and H. W. Werner, eds. "Secondary Ion Mass Spectrometry: SIMS XIII, (Nara)." *Applied Surface Science*, Vol. 203–204. North-Holland: Elsevier, 2003.
205. Benninghoven, A., P. Bertrand, H.-N.Migeon, and H. W. Werner, eds. *Secondary Ion Mass Spectrometry: SIMS XII, (Brussels)*. Amsterdam: Elsevier, 2000.
206. Gillen, G., R. Lareau, J. Bennett, and F. Stevie, eds. *Secondary Ion Mass Spectrometry: SIMS XI, (Orlando)*. New York: Wiley, 1998.
207. Wilson, R. G., F. A. Stevie, and C. A. Magee. *Secondary Ion Mass Spectrometry: A Practical Handbook for Depth Profiling and Bulk Impurity Analysis*. New York: Wiley, 1989.
208. Schumacher, M., H. N. Migeon, and B. Rasser. *Proc. SIMS VIII Conf. Amsterdam* (1991): 49.
209. Janssens, T., and W. Vandervorst. *Proc. SIMS XII Conf. Brussels* (1999): 151.
210. Maul, J., F. Schultz, and K. Wittmaack. *Rad. Effects* 18 (1973): 211.
211. McHugh, J. A. *Rad. Effects* 21 (1974): 209.
212. Williams, P. *Proc. SIMS XI, Orlando* (1997): 3.
213. Mayer, J. W., B. Y. Tsaur, S. S. Lau, and L-S. Hung. *Nucl. Instrum. Methods* B182/183 (1981): 1.
214. Littmark, U., and W. O. Hofer. *Nucl. Instrum. Methods* 168 (1981): 329.
215. Wittmaack, K. *Vacuum* 34 (1984): 119.
216. SRIM-2000, J. Ziegler.
217. Vickerman, J. C., and D. Briggs. *ToF SIMS: Surface Analysis by Mass Spectrometry*. Manchester, NH: IM Publications and Surface Spectra Ltd, 2001.
218. Vandervorst, W., et al. "Errors in Near-Surface and Interfacial Profiling of Boron and Arsenic." *Appl. Surf. Sci.* 231–232 (2004): 618.
219. Jiang, Z., et al. *Appl. Phys. Lett.* 73 (1998): 315.

220. Wittmaack, et al. *J. Vac. Sci. Technol.* B16 (1998): 272.
221. Several articles on sputter rate transients in: *J. Vac. Sci. Tech.* B18 (2000). Schueler, B., and D. F. Reich. *J. Vac. Sci. Tech.* B18 (2000): 496; Cooke, G. A., T. J. Ormsby, M. G. Dowsett, C. Parry, A. Murell, and E. J. H. Collard. *J. Vac. Sci. Tech.* B18 (2000): 493 Wittmack, K., J. Griesche, H. J. Osten, and S. B. Patel. *J. Vac. Sci. Tech.* B18 (2000): 524; Ronsheim, P. A., and J. J. Murphy. *J. Vac. Sci. Tech.* B18 (2000): 501.
222. Williams, P., and J. E. Baker. *Nucl. Instrum. Methods* 182/183 (1981): 15.
223. Boudewijn, P. R., H. W. P. Akerboom, and N. M. C. Kempeners. *Spectrochim. Acta* 39B (1984): 1567.
224. Hues, S. M., and P. Williams. *Nucl. Instrum. Methods* B15 (1986): 206.
225. Bullis, W. M. "Oxygen Concentration Measurement." In *Oxygen in Silicon*, edited by F. Shimura, 94. Boston, MA: Academic Press, 1994.
226. Shaffner, T. J., and D. K. Schroder. "Characterization Techniques for Oxygen in Silicon." In *Oxygen in Silicon*, edited by F. Shimura, 53–93. New York: Academic Press, 1994.
227. Pajot, B. "Characterization of Oxygen in Silicon by Infrared Absorption." *Analysis* 5 (1977): 32–42.
228. Bullis, W. M., S. Perkowitz, and D. G. Seiler. *Survey of Optical Characterization Methods for Materials, Processing, and Manufacturing in the Semiconductor Industry*. Washington, DC: National Institute of Standards and Technology, 1995 (NIST 400-98).
229. Perkowitz, S., D. G. Seiler, and W. M. Duncan. "Optical Characterization in Microelectronics Manufacturing." *J. Res. NIST* 99 (1994): 605–39.
230. Baber, S. C. "Net and Total Shallow Impurity Analysis of Silicon by Low Temperature Fourier Transform Infrared Spectroscopy." *Thin Solid Films* 72 (1980): 201–10.
231. Wagner, P. "Infrared Absorption of Interstitial Oxygen in Silicon at Low Temperatures." *Appl. Phys. A* 53 (1991): 20–5.
232. Moore, C. J. L., and C. J. Miner. "A Spatially Resolved Spectrally Resolved Photoluminescence Mapping System." *J. Cryst. Growth* 103 (1990): 21–7.
233. Tajima, M., T. Masui, D. Itoh, and T. Nishino. "Calibration of the Photoluminescence Method for Determining As and Al Concentrations in Si." *J. Electrochem. Soc.* 137 (1990): 3544–51.
234. Duncan, W. M., and M. L. Eastwood. "Fourier Transform Photoluminescence Analysis of Semiconductor Materials." *Proc. SPIE* 822 (1987): 172–80.
235. Thewalt, M. L. W., A. G. Steele, and J. E. Huffman. "Photoluminescence Studies of Ultrahigh-Purity Epitaxial Silicon." *Appl. Phys. Lett.* 49 (1986): 1444–6.
236. Tajima, M. "Determination of Boron and Phosphorus Concentration in Silicon by Photoluminescence Analysis." *Appl. Phys. Lett.* 32 (1978): 719–21.
237. Wolfe, J. P., and A. Mysyrowicz. "Excitonic Matter." *Sci. Am.* 250 (1984): 98–107.
238. Smith, K. K. "Photoluminescence of Semiconductor Materials." *Thin Solid Films* 84 (1981): 171–82.
239. Murray, R., K. Graff, B. Pajot, K. Strijckmans, S. Vandendriessche, B. Griepink, and H. Marchandise. "Interlaboratory Determination of Oxygen in Silicon for Certified Reference Materials." *J. Electrochem. Soc.* 139 (1992): 3582–6.
240. Baghdadi, A., W. M. Bullis, M. C. Croarkin, Y. Li, R. I. Scace, R. W. Series, P. Stallhofer, and M. Watanabe. "Interlaboratory Determination of the Calibration Factor for the Measurement of the Interstitial Oxygen Content of Silicon by Infrared Absorption." *J. Electrochem. Soc.* 136 (1989): 2015–24.
241. "Standard Test Methods for Oxygen Precipitation Characterization of Silicon Wafers by Measurement of Interstitial Oxygen Reduction." *ASTM F1239*, 210–1. American Society for Testing and Materials, 1989.
242. Vandendriessche, S., B. Griepink, H. Marchandise, B. Pajot, R. Murray, K. Graff, and K. Strijckmans. *The Certification of a Reference Material for the Determination of Oxygen in Semiconductor Silicon by Infrared Spectrometry*. CRM 369 CA: 115(10)104911g, American Chemical Society, 1992, p. 339.
243. Gladden, W. K., S. R. Slaughter, W. M. Duncan, and A. Baghdadi. *Automatic Determination of the Interstitial Oxygen Content of Silicon Wafers Polished on Both Sides*. Washington, DC: National Institute of Standards and Technology, 1988 (NIST 400-81).

244. Becker, D. A., R. M. Lindstrom, and T. Z. Hossain. "International Intercomparison for Trace Elements in Silicon Semiconductor Wafers by Neutron Activation Analysis." In *Semiconductor Characterization: Present Status and Future Needs*, edited by W. M. Bullis, D. G. Seiler, and A. C. Diebold, 335–41. Woodbury, NY: American Institute of Physics, 1996.
245. McGuire, S. C., T. Z. Hossain, A. J. Filo, C. C. Swanson, and J. P. Lavine. "Neutron Activation for Semiconductor Material Characterization." In *Semiconductor Characterization: Present Status and Future Needs*, edited by W. M. Bullis, D. G. Seiler, and A. C. Diebold, 329–34. Woodbury, NY: American Institute of Physics, 1996.
246. Smith, A. R., R. J. McDonald, H. Manini, D. L. Hurley, E. B. Norman, M. C. Vella, and R. W. Odom. "Low-Background Instrumental Neutron Activation Analysis of Silicon Semiconductor Materials." *J. Electrochem. Soc.* 143 (1996): 339–46.
247. Paul, R. L., and R. M. Lindstrom. "Applications of Cold Neutron Prompt Gamma Activation Analysis to Characterization of Semiconductors." In *Seimiconductor Characterization: Present Status and Future Needs*, edited by W. M. Bullis, D. G. Seiler, and A. C. Diebold, 342–5. Woodbury, NY: American Institute of Physics, 1996.
248. Lindstrom, R. M. "Activation Analysis of Electronics Materials." In *Microelectronics Processing: Inorganic Materials Characterization*, edited by L. A. Casper, 294–307. Washington, DC: American Chemical Society, 1986.
249. Blondiaux, G., J-L. Debrun, and C. J. Maggiore. "Charged Particle Activation Analysis." In *Handbook of Modern Ion Beam Materials Analysis*, edited by J. R. Tesmer, M. Nastasi, J. C. Barbour, C. J. Maggiore, and J. W. Mayer, 205–30. Pittsburgh, PA: Materials Research Society, 1995.
250. Haas, E. W., and R. Hofmann. "The Application of Radioanalytical Methods in Semiconductor Technology." *Solid-State Electron.* 30 (1987): 329–37.
251. Hirvonen, J-P. "Nuclear Reaction Analysis: Particle-Gamma Reactions." In *Handbook of Modern Ion Beam Materials Analysis*, edited by J. R. Tesmer, M. Nastasi, J. C. Barbour, C. J. Maggiore, and J. W. Mayer, 167–92. Pittsburgh, PA: Materials Research Society, 1995.
252. Lanford, W. A. "Nuclear Reactions for Hydrogen Analysis." In *Handbook of Modern Ion Beam Materials Analysis*, edited by J. R. Tesmer, M. Nastasi, J. C. Barbour, C. J. Maggiore, and J. W. Mayer, 193–204. Pittsburgh, PA: Materials Research Society, 1995.
253. Downing, R. G., and G. P. Lamaze. "Nondestructive Characterization of Semiconductor Materials Using Neutron Depth Profiling." In *Semiconductor Characterization: Present Status and Future Needs*, edited by W. M. Bullis, D. G. Seiler, and A. C. Diebold, 346–50. Woodbury, NY: American Institute of Physics, 1996.
254. Downing, R. G., J. T. Maki, and R. F. Fleming. "Application of Neutron Depth Profiling to Microelectronic Materials Processing." In *Microelectronics Processing: Inorganic Materials Characterization*, edited by L. A. Casper, 163–80. Washington, DC: American Chemical Society, 1986.
255. Ehrstein, J. R., R. G. Downing, B. R. Stallard, D. S. Simons, and R. F. Fleming. "Comparison of Depth Profiling 10B in Silicon Using Spreading Resistance Profiling, Secondary Ion Mass Spectrometry, and Neutron Depth Profiling." In *Semiconductor Processing, ASTM STP 850*, edited by D. C. Gupta, 409–25. Philadelphia, PA: American Society for Testing and Materials, 1984.
256. Ricci, E., and R. L. Hahn. "Rapid Calculation of Sensitivities, Interferences, and Optimum Bombarding Energies in ^{3}He Activation Analysis." *Anal. Chem.* 40 (1968): 54.
257. Feldman, L. C., and J. W. Mayer. *Fundamentals of Surface and Thin Film Analysis*. New York: North-Holland, 1986.
258. Leavitt, J. A., L. C. McIntyre Jr.., and M. R. Weller. "Backscattering Spectrometry." In *Handbook of Modern Ion Beam Materials Analysis*, edited by J. R. Tesmer, and M. Nastasi, 37–81. Pittsburgh, PA: Materials Research Society, 1995.
259. Anthony, J. M. "Ion Beam Characterization of Semiconductors." In *Materials Characterization: Present Status and Future Needs*, edited by W. M. Bullis, D. G. Seiler, and A. C. Diebold, 366–76. Woodbury, NY: American Institute of Physics, 1996.
260. Jacobsen, F. M., M. J. Zarcone, D. Steski, K. Smith, P. Thieberger, K. G. Lynn, J. Throwe, and M. Cholewa. "Detection of Heavy Trace Impurities in Silicon." *Semicond. Int.* (1996): 243–8.

261. Mendenhall, M. H., and R. A. Weller. "High-Resolution Medium-Energy Backscattering Spectrometry." *Nucl. Instrum. Methods B* 59/60 (1991): 120–3.
262. Baglin, J. E. E. "Elastic Recoil Spectrometry." In *Encyclopedia of Materials Characterization*, edited by C. R. Brundle, C. A. Evans Jr.., and S. Wilson, 488–501. Boston, MA: Butterworth-Heinemann, 1992.
263. Barbour, J. C., and B. L. Doyle. "Elastic Recoil Detection: ERD." In *Handbook of Modern Ion Beam Materials Analysis*, edited by J. R. Tesmer, and M. Nastasi, 83–138. Pittsburgh, PA: Materials Research Society, 1995.
264. Musket, R. G. "Particle-Induced X-Ray Emission." In *Encyclopedia of Materials Characterization*, edited by C. R. Brundle, C. A. Evans Jr.., and S. Wilson, 357–69. Boston, MA: Butterworth-Heinemann, 1992.
265. Tabacniks, M. H., A. J. Kellock, and J. E. E. Gaglin. "PIXE for Thin Film Analysis." In *Application of Accelerators in Research and Industry: Proceedings of the Fourteenth International Conference*, edited by J. L. Duggan, and I. L. Morgan, 563–6. Woodbury, NY: American Institute of Physics, 1997.
266. Swanson, M. L. "Channeling." In *Handbook of Modern Ion Beam Materials Analysis*, edited by J. R. Tesmer, and M. Nastasi, 231–300. Pittsburgh, PA: Materials Research Society, 1995.
267. Evans Analytical Group. Rutherford Backscattering Spectroscopy Theory Tutorial. http://www.eaglabs.com/en-US/references/tutorial/rbstheo/chanling.html (accessed on February 21, 2007).
268. Morris, W. G., S. Fesseha, and H. Bakhru. "Microbeam RBS and PIXE Applied to Microelectronics." *Nucl. Instrum. Methods B* 24/25 (1987): 635–7.
269. Cross, B. J., D. C. Wherry, and T. H. Briggs. "New Methods for High-Performance X-Ray Fluorescence Thickness Measurements." *Plating Surf. Finishing* (1988): 1–7.
270. Parekh, N., C. Nieuwenhuizen, J. Borstrok, and O. Elgersma. "Analysis of Thin Films in Silicon Integrated Circuit Technology by X-Ray Fluorescence Spectrometry." *J. Electrochem. Soc.* 138 (1991): 1460–5.
271. Ernst, S., C. Lee, and J. Lee. "Thickness Measurement of Aluminum, Titanium, Titanium Silicide, and Tungsten Silicide Films by X-Ray Fluorescence." *J. Electrochem. Soc.* 135 (1988): 2111–3.
272. Eichinger, P. "Total Reflection X-Ray Fluorescence." In *Encyclopedia of Materials Characterization*, edited by C. R. Brundle, C. A. Evans Jr., S. Wilson, and L. E. Fitzpatrick, 349–56. Boston, MA: Butterworth-Heinemann, 1992.
273. Diebold, A. C., and B. Doris. "A Survey of Non-Destructive Surface Characterization Methods Used to Insure Reliable Gate Oxide Fabrication for Silicon IC Devices." *Surf. Interface Anal.* 20 (1993): 127–39.
274. Klockenkämper, R. *Total-Reflection X-Ray Fluorescence Analysis*. New York: Wiley, 1997.
275. Jenkins, R., R. W. Gould, and D. Gedcke. *Quantitative X-Ray Spectrometry*. New York: Marcel Dekker, Inc., 1997.
276. Nichols, M. C., D. R. Boehme, R. W. Ryon, D. Wherry, B. Cross, and G. Aden. "Parameters Affecting X-Ray Microfluorescence (XRMF) Analysis." In *Advances in X-Ray Analysis*, edited by C. S. Barrett, J. V. Gilfrich, R. Jenkins, D. E. Leyden, J. C. Russ, and P. K. Predecki, 45–51. New York: Plenum Publishing Corporation, 1987.
277. Isaacs, E. D., K. Evans-Lutterodt, M. A. Marcus, A. A. Macdowell, W. Lehnert, J. M. Vandenberg, S. Sputz., et al. "X-Ray Microbeam Techniques and Applications." In *Diagnostic Techniques for Semiconductor Materials and Devices*, edited by P. Rai-Choudhury, J. L. Benton, D. K. Schroder, and T. J. Shaffner, 49–67. Pennington, NJ: The Electrochemical Society, 1997.
278. Attaelmanan, A., P. Voglis, A. Rindby, S. Larsson, and P. Engstrom. "Improved Capillary Optics Applied to Microbeam X-Ray Fluorescence: Resolution and Sensitivity." *Rev. Sci. Instrum.* 66 (1995): 24–7.
279. Brundle, C. R. "X-Ray Photoelectron Spectroscopy." In *Encyclopedia of Materials Characterization*, edited by C. R. Brundle, C. A. Evans Jr., S. Wilson, and L. E. Fitzpatrick, 282–99. Boston, MA: Butterworth-Heinemann, 1992.
280. Nieveen, W. In *Proceedings 207th ECS Meeting*, Vol. PV 2005-01, 208–22, Quebec, Canada 2005.
281. Seah, M. P., and W. A. Dench. "Quantitative Electron Spectroscopy of Surfaces: A Standard Data Base for Electron Inelastic Mean Free Paths in Solids." *Surf. Interface Anal.* 1 (1979): 2–11.

282. Moulder, J. F., W. F. Stickle, P. E. Sobol, and K. D. Bomben. *Handbook of X-Ray Photoelectron Spectroscopy*. Eden Prairie, MN: Physical Electronics Inc., 1995.
283. Moulder, J., ed. *The PHI Interface*. Vol. 21 (1), 6. Chanhassen, MN: Physical Electronics, 2005.
284. Strausser, Y. E. "Auger Electron Spectroscopy." In *Encyclopedia of Materials Characterization*, edited by C. R. Brundle, C. A. Evans Jr., S. Wilson, and L. E. Fitzpatrick, 310–23. Boston, MA: Butterworth-Heinemann, 1992.
285. Auger, M. P., and M. J. Perrin. "Sur les Rayons β Secondaires Produits dans un Gaz par des Rayons X." *Comptes rendus* 180 (1925): 65–8 (orig.).
286. Briggs, D., and M. P. Seah. *Practical Surface Analysis by Auger and X-Ray Photoelectron Spectroscopy*. New York: Wiley, 1984.
287. Harris, L. A. "Analysis of Materials by Electron-Excited Auger Electrons." *J. Appl. Phys.* 39 (1968): 1419–27.
288. *Handbook of Auger Electron Spectroscopy*. Eden Prairie, MN: Physical Electronics Industries, Inc. 1976.
289. Shaffner, T. J. "Rapid Semi-Quantitative Analysis for Routine Applications of Scanning Auger Microscopy." *Scanning Electron Microsc.* I (1980): 479–86.
290. Hall, P. M., and J. M. Morabito. "Compositional Depth Profiling by Auger Electron Spectroscopy." *CRC Crit. Solid State Mater. Sci.* 8 (1978): 53–67.
291. Hofmann, S., and A. Zalar. "Auger Electron Spectroscopy Depth Profiling of Ni/Cr Multilayers by Sputtering with N_2^+ Ions." *Thin Solid Films* 60 (1979): 201–11.
292. McCarthy, G. J., J. M. Holzer, W. M. Syvinski, K. J. Martin, and R. G. Garvey. "Evaluation of Reference X-Ray Diffraction Patterns in the ICDD Powder Diffraction File." In *Advances in X-Ray Analysis*, edited by C. S. Barrett, J. V. Gilfrich, I. C. Noyan, T. C. Huang, and P. K. Predecki, 369–76. New York: Plenum Press, 1991.
293. Goehner, R. P., and M. C. Nichols. "X-Ray Powder Diffraction." In *Metals Handbook Ninth Edition: Volume 10 Materials Characterization*, edited by R. E. Whan, 333–43. Metals Park, OH: American Society for Metals, 1986.
294. Toney, M. F. "X-Ray Diffraction." In *Encyclopedia of Materials Characterization*, edited by R. C. Brundle, C. A. J. Evans, and S. Wilson, 198–213. Boston, MA: Butterworth-Heinemann, 1992.
295. Cullity, B. D. *Elements of X-Ray Diffraction*. Reading, MA: Addison-Wesley Publishing Co., 1978.
296. Adams, B. L. "Crystallographic Texture Measurement and Analysis." In *Metals Handbook Ninth Edition: Volume 10 Materials Characterization*, edited by R. E. Whan, 357–79. Metals Park, OH: American Society for Metals, 1986.
297. Bowen, D. K., and B. K. Tanner. *High Resolution X-Ray Diffractometry and Topography*. London, U.K.: Taylor & Francis Publishers Ltd, 1998.
298. Hart, M. "Bragg Angle Measurement and Mapping." *J. Cryst. Growth* 55 (1981): 409–27.
299. Tanner, B. K. "X-Ray Scattering for Semiconductor Characterization." In *Semiconductor Characterization: Present Status and Future Needs*, edited by W. M. Bullis, D. G. Seiler, and A. C. Diebold, 263–72. New York: American Institute of Physics, 1996.
300. Patel, J. R. "X-Ray Anomalous Transmission and Topography of Oxygen Precipitation in Silicon." *J. Appl. Phys.* 44 (1973): 3903–6.
301. Pangborn, R. N. "X-Ray Topography." In *Metals Handbook Ninth Edition: Volume 10 Materials Characterization*, edited by R. E. Whan, 365–79. Metals Park, OH: American Society for Metals, 1986.
302. Patel, J. R. "X-Ray Diffuse Scattering from Silicon Containing Oxygen Clusters." *J. Appl. Cryst.* 8 (1975): 186–91.
303. Koppel, L. N., and L. Parobek. "Thin-Film Metrology by Rapid X-Ray Reflectometry." In *International Conference on Characterization and Metrology for ULSI Technology*, edited by D. G. Seiler, 1. New York: American Institute of Physics, 1998.
304. Miyauchi, A., K. Usami, and T. Suzuki. "X-Ray Reflectivity Measurement of an Interface Layer between a Low Temperature Silicon Epitaxial Layer and HF-Treated Silicon Substrate." *J. Electrochem. Soc.* 141 (1994): 1370–4.

305. Padmore, H. A., and P. Pianetta. "X-Ray Microscopy and TXRF: Emerging Synchrotron Techniques for Semiconductor Characterization." In *International Conference on Characterization and Metrology for ULSI Technology*, edited by D. G. Seiler, New York: American Institute of Physics, 1998.
306. Qadri, S. B., D. Ma, and M. Peckerar. "Double-Crystal X-Ray Topographic Determination of Local Strain in Metal-Oxide-Semiconductor Device Structures." *Appl. Phys. Lett.* 51 (1987): 1827–9.
307. Zaumseil, P., U. Winter, M. Servidori, and F. Cembali. "Determination of Defect and Strain Distribution in Ion Implanted and Annealed Silicon by X-Ray Triple Crystal Diffractometry." In *Gettering and Defect Engineering in Semiconductor Technology/GADEST 1987*, edited by H. Richter, 195–9. Germany: Academy Sciences, 1987.
308. Kawado, S., S. Kojima, and I. Maekawa. "Influence of Crystal Imperfection on High-Resolution Diffraction Profiles of Silicon Single Crystals Measured by Highly Collimated X-Ray Beams." *Appl. Phys. Lett.* 58 (1991): 2246–8.
309. Huff, H. R., H. F. Schaake, J. T. Robinson, S. C. Baber, and D. Wong. "Some Observations on Oxygen Precipitation/Gettering in Device Processed Czochralski Silicon." *J. Electrochem. Soc.* 130 (1983): 1551–5.
310. Tuomi, T., M. Tuominen, E. Prieur, J. Partanen, J. Lahtinen, and J. Laakkonen. "Synchrotron Section Topographic Study of Czochralski-Grown Silicon Wafers for Advanced Memory Circuits." *J. Electrochem. Soc.* 142 (1995): 1699–701.
311. Partanen, J., T. Tuomi, and K. Katayama. "Comparison of Defect Images and Density between Synchrotron Section Topography and Infrared Light Scattering Microscopy in Heat-Treated Czochralski Silicon Crystals." *J. Electrochem. Soc.* 139 (1992): 599–604.
312. Jiang, B. L., F. Shimura, and G. A. Rozgonyi. "X-Ray Moire Pattern in Dislocation-Free Silicon-on-Insulator Wafers Prepared by Oxygen Ion Implantation." *Appl. Phys. Lett.* 56 (1990): 352–4.
313. Turrell, G., and J. Corset. *Raman Microscopy, Developments and Applications*. London, U.K.: Academic Press Limited, 1996.
314. Ferraro, J. R., and K. Nakamoto. *Introductory Raman Spectroscopy*. San Diego, CA: Academic Press, Inc., 1994.

29

Failure Analysis

29.1 Introduction **29**-1

29.2 Failure Site Isolation **29**-4
Electrical Characterization • Tools for Bench Electrical Characterization • Die Exposure Techniques • Global Techniques • Probing

29.3 Physical Analysis Tools **29**-14
Package Analysis • Deprocessing • Parallel Polishing • Cross-Section Analysis • Microscopy • Transmission Electron Microscopy

29.4 Chemical Characterization **29**-18
X-Ray Analysis (Energy Dispersive) • Auger • Secondary Ion Mass Spectroscopy (SIMS) • Microspot FTIR • Others

29.5 Future of Failure Analysis **29**-21

References **29**-22

Lawrence C. Wagner

Texas Instruments, Inc.

29.1 Introduction

Failure analysis (FA) describes the process of diagnosis of defective integrated circuits (ICs) [1]. Historically, this has been most closely tied to the analysis of packaged devices, primarily customer returns and qualification failures. However, defective circuits that require analysis occur at all stages of manufacture and use.

Some of the primary applications include design debug, product and process ramps, yield enhancement, reliability test failures, and customer returns (see Table 29.1). There is a fairly common tool set for these diagnostic activities although the implementations may vary significantly.

In general, the FA process consists of two phases (see general FA process flow in Figure 29.1). The first is determining the electrical cause of failure or failure site isolation. This is the process of narrowing the scope of analysis from a complex integrated circuit down to a much simpler problem, the analysis of a single failing net, transistor, thin film, or junction. The tools used in the isolation process begin with an electrical test. In fact, electrical testing can in some cases provide failure site isolation by itself. This is particularly true in case of single bit memory failures where the failure site is quickly isolated through electrical test alone to a very small area. The temperature and voltage dependence of the single bit failure can further isolate the failure to a particular structure in the memory bit [2]. Logic failures can also be isolated or partially isolated by electrical testing. This is particularly true of IC's with scan based Design for Test structures (DFT) [3,4]. Electrical testing also provides a method of placing the device in a failing condition. This failing condition is required for the application of the various physical failure site isolation tools. These tools can be broadly categorized

TABLE 29.1 Failure Analysis (FA) Is a Loosely Defined Term

Diagnostic Activity	Primary Failure Site Isolation Techniques	Primary Physical Analysis	Comments
Design debug	Mechanical probe E-beam probe	Circuit analysis	Focused on identification of design errors or marginality
Wafer fab yield improvement: defect based	Electrical data analysis overlaid on defect map Global techniques	Deprocessing Focused ion beam (FIB) cross section	Statistical approach with emphasis on efficient analysis of a large sample size
Wafer fab yield improvement: parametric based	Electrical analysis of test structures	Cross section of test structures	Interest is primarily in improvement in parameter means and distribution
Wafer fab yield improvement: unmodeled loss	Mechanical probe E-beam probe Emission microscopy	Circuit analysis Deprocessing Cross section	Typically associated with a design feature which interacts with the wafer fab process for a higher defect density
Assembly test yield improvement	Non-destructive package analysis: x-ray, scanning acoustic microscopy, time domain reflectometry	Decapsulation visual/ scanning electron microscope inspection	Statistically based on most common failure bins
Qualification failures	All	All	Typically unique failures must be successfully analyzed No statistical population
Customer returns	All	All	Sample sizes vary greatly and statistical significance varies

Several diagnostic activities that are commonly referred to as FA are summarized in Table 29.1. Failure analysis can have significantly different connotations in various semiconductor environments. They share a common tool set with varying emphasis and application.

as global and probing techniques. Global techniques use secondary events such as thermal or light generation from a failure site to isolate a failure. These techniques are particularly powerful because they can be performed quickly and do not require an understanding of the circuit operation. Probing techniques allow direct electric measurements on electrical nets or nodes in the failed device. Localization of defects by probing is frequently tedious and time consuming. The failure site isolation process generally leads to an electrical cause of failure, for example a net shorted to ground or an open net. It may also provide an accurate location of the failure, for example the precise location of an open conductor or shunted signal lines.

Once isolated, the physical cause of the failure remains to be determined. This is achieved through physical observation of the defect. Additional electrical characterization may be required during the process of exposing the defective area to further narrow down the location and type of failure. When a particle or other contamination is involved, chemical analysis also forms a critical part of understanding the source of the contamination. However, the physical cause of failure can cover a wide range of possible failure mechanisms. In a design debug activity, the physical cause may be a layout anomaly. In a package related failure, it could be an adhesion failure between the mold compound and die surface.

The task of physical analysis utilizes a broad set of tools, which include various forms of microscopy as well as chemical analysis. Understanding the failure may entail other tasks such as circuit simulations to verify that the physical cause of failure matches the electrical cause of failure. This is particularly true for design debug activities. This is also often the case for subtle analog circuit failures where small process shifts or circuit element mismatches can play a critical role.

In a broader sense, the FA process described above is an element of improvement processes in the semiconductor industry (see Figure 29.2). The first part of this process is determining which failures are

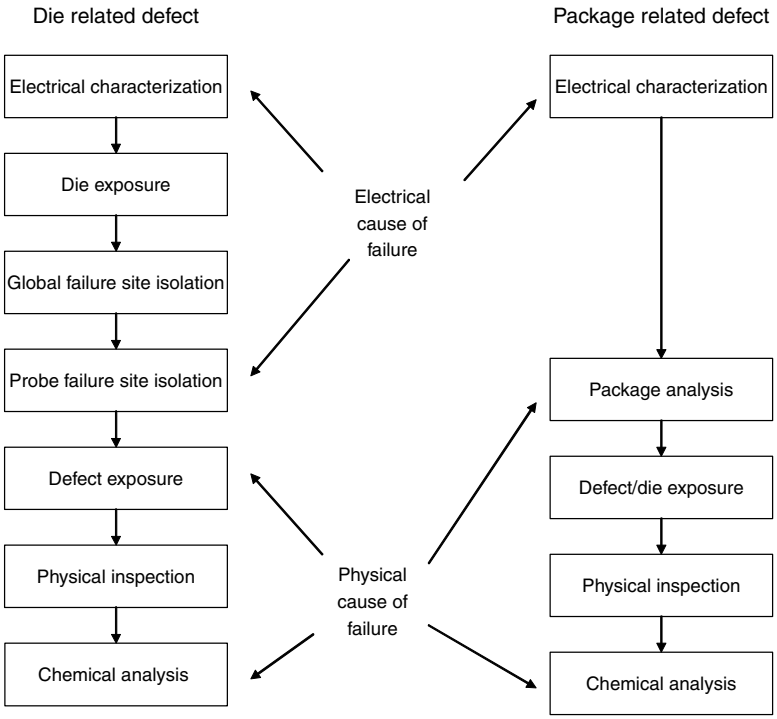


FIGURE 29.1 The general process flow of failure analysis (FA) can be broken down in the determination of electrical cause of failure followed by determination of the physical cause of failure.

most significant to analyze and understand. The failure modes can provide signatures [5,6] of various categories of failures to sample from. A good example is yield analysis, where Pareto distributions of failure modes, electrical signatures of failure, drive the selection of devices to analyze. Time is expended in analysis of the most common failures because elimination of these failures will have the greatest potential impact on yield.

Closure of the improvement process consists of root cause identification and corrective actions. The root cause of the failure goes beyond understanding the physical cause of the failure. A good example is the conductive particle creating a short circuit. The physical cause of this failure can be fully understood in terms of the resistance, composition, and location of the particle in the FA process. The root cause of this failure additionally requires an understanding of how the particle was generated in the wafer processing tool. For example, the root cause of the particle might be mechanical wear in a load lock to the wafer fab tool, which is generating particles from a particular component in the load lock with the same composition as the particle.

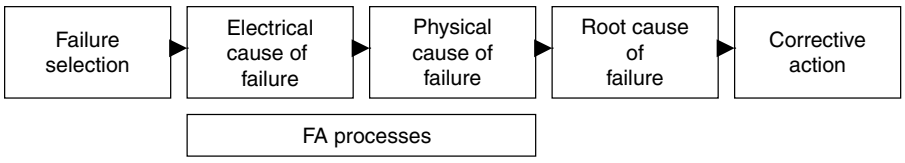


FIGURE 29.2 General flow of diagnosis in semiconductor is shown. FA generally includes determining the electrical and the physical cause of failure.

29.2 Failure Site Isolation

As discussed above, failure site isolation begins with an electrical characterization [7]. Understanding how a device fails electrically is a critical first step in the failure site isolation flow. In some cases, it can provide a relatively complete understanding of the failure e.g., single bit memory failures. In most cases, it provides a good understanding of how to proceed with physical failure site isolation. It also provides an electrical test set-up to place the device into a failing condition. This is essential for the various techniques used in the physical failure isolation process. In addition, die exposure is a requirement to proceed to the physical failure site isolation except for the analysis of unpackaged devices or wafers. Global techniques are typically the first set of techniques applied, since they provide quick isolations when successful. If the global techniques are unsuccessful, more time consuming probing techniques must be employed. In some cases, a more detailed understanding of the component level failure is obtained by isolating and characterizing the failed electrical component.

29.2.1 Electrical Characterization

Electrical characterization begins with the analysis of the electrical failure signature. In many cases, this is the datalog from automatic test equipment (ATE) data. If ATE is not readily available, other methods may be attempted before shipping a part to a test floor. These include curve tracing, system functionality testing, (typically a populated circuit board with a socket in place of the component which is to be tested and used to very system operation) and emulation boards. A datalog is simply an electronic output of the results from various tests performed as a part of the production test program. Where a large number of datalogs are to be examined such as in wafer fab yield analysis, it is customary to use binning (putting failures into categories by type of electrical test failed) to provide a first pass sorting of the devices. This helps to determine which devices to focus analysis on. In addition, this binning may be combined with wafer mapping to identify patterns within the wafer lots that may also drive the direction of the analysis. As die identification becomes more common, it becomes realistic to apply the concepts of wafer mapping into reliability and customer return failures [8].

The interpretation of datalogs is a key first step in the FA process. In general, the focus is on a hierarchy of failure modes. This hierarchy is used to determine the simplest electrical test conditions that can be used to create the failing condition. The first level in this hierarchy is continuity failures, opens and shorts. An open or shorted pin usually generates a large number of parametric and functional failures. Since the parametric and functional failures are due to the open or short pin, the only electrical characteristic, which must be analyzed, is the continuity failure. For example, an output shorted to ground will result in parametric failures such as voltage output high (VOH) as well as functional failures at each vector where the shorted output is expected to be high. These tend to drive quick isolation since the continuity testing is typically a measure of the assembly interconnections and diodes directly connected to the pin. Hence the failure must occur within the pin circuitry or in the package connection to the pin.

The second level of the failure hierarchy is parametric failures. There are a number of parameters that are measured for each input and for each output of a circuit during production electrical testing. Many of these parameters can be measured without regard to the bias of other pins except power and ground pins. However, measurements of some output pin parameters can require biasing that output into a particular logic state in order to make the measurement. In such cases, an output parameter failure may represent only a failure to get the output into the correct state for the measurement and the failure, which must be analyzed and may in fact be a functional failure preventing the output from being in the anticipated state. However most input or output parametric failures are generally isolated to the circuitry directly attached to the failing pin.

In addition to input and output parameters, power supply leakage parameters are critical for FA. While the leakage between the power supplies can typically occur in any area of the die, these leakage

failures are particularly amenable to the global isolation techniques discussed below. Functional failures without power supply leakages, on the other hand, are typically isolated using scan design for test methods or must be probed for isolation. In complementary metal-oxide-silicon (CMOS) circuits, quiescent leakage current (IDDQ) leakage is typically measured in states where a low leakage level is anticipated for a good device. This IDDQ leakage may be state dependent i.e., it may only occur at specific logical conditions or it may be state independent. For state independent leakage, electrical biasing for isolation is not as critical. However, for state dependent leakages, the capability to put the device into the failing condition is essential for successful analysis. Essentially, the device must be electrically preconditioned into the failing state in order to measure or isolate the IDDQ failure. This can require sequentially entering a large number of input patterns. Static CMOS logic typically has an extremely low IDDQ, which allows detection of very low level leakages in the device logic. For device without low leakage states, power supply leakage may be a useful parameter for FA even if it is not measured as part of the production test program. By comparing the leakage on a failing device to that on a known good device, it may be possible to isolate the source of the increased leakage which may be causing the observed failure. Variations in IDDQ with temperature can also provide useful insights into the type of failure mechanism being observed [9].

The final category of failures is functional failures. These are failures that do not exhibit abnormal IDDQ leakages or the other parametric anomalies. Certain categories of functional failures can be isolated with good precision in the IC. This includes memory failures for which electrical data may precisely identify an area for physical analysis. This also includes scan failures for which tools exist to isolate scan failures to certain nets, in some case as precisely as a single net in the circuit. With the exception of the categories identified above, function failures are typically the most difficult type of failure to analyze. This is especially true of frequency dependent and temperature dependent functional failures. Functional FA is particularly demanding because global techniques tend to be less successful on this category of failures, forcing the analyst to resort to probing approaches.

29.2.2 Tools for Bench Electrical Characterization

Curve tracer [10] and parametric analyzer I–V curves are typical adequate to understand continuity failures. While the datalog will identify the failing pins, I–V curves provide a better understanding of the failure. For example, a high resistance on a pin can be viewed as open on a datalog. Parametric FA requirements range from simple leakage failures, which can be observed with a curve tracer or parametric analyzer to more difficult parameters, which require control of various pins on a device. This typically occurs to put an output into a particular logic state. For example, measurement of VOH and current, output high (IOH) requires putting the output into a high state. Depending on the device type, this may require control of at least one and usually more input pins.

One of the factors, which contribute to the increased difficulty of FA, is the increase in device complexity. In terms of the electrical characterization, this is reflected in the increase in typical pin counts as well as transistor count. The tools used for the electrical characterization of devices and biasing for failure site isolation depends very heavily on the complexity of the devices (See Table 29.2).

For low pin count devices (2–28 pins), electrical characterization is best performed with simple bench test electronics. For continuity failures, this includes curve tracers and parametric analyzers. For parametric failures, this includes a range of meters and power supplies as well as customized test boxes, which are frequently hand built. Powered curve tracing, I–V curves generated with a static bias on the pins provides a great deal of information about the possible location of the failure in the input or output structure. Analysis of functional failures usually also requires simple function or pattern generation capability as well as oscilloscopes for output measurement.

Medium pin count devices through 300 pins typically require a different range of equipment. The management and measurement of a larger number of pins require a level of automation for efficiency. Typically a pin matrix with an automated curve tracer is significantly more efficient for continuity failures

TABLE 29.2 Various Types of Electronics Are Used for Electrical Characterization of Failed Devices

	I–V Characterization	Parametric Analysis	Functional Testing/Timing
Low pin count (2–28 pins)	Standard curve tracer Parameter analyzer	Power supply Meters Parameter analyzers Custom boxes	Pulse/pattern generators Oscilloscopes
Medium pin count (28–300 pins)	Pin matrices Automated curve tracer	Pin matrices Automated curve tracer Power supplies Meters	High pin count pattern generators Logic analyzers Application specific integrated circuit verification testers
High pin count (>300 pins)	Automated curve tracer	Automated curve tracers ATE	Automatic test equipment (ATE)
Ultra high pin count (>1000 pins)	Automated curve tracers ATE	Automated curve tracers ATE	ATE

A limited summary of the most common tools is presented above.

than manual curve tracers. Parametric analysis may still be performed on a bench with a pin matrix replacing the custom boxes. Some of the parametric analysis can also be performed on the automated curve tracers. Functional FA requires a much higher level of pin control and simple bench electronics are increasingly ineffective as the pin count increases through this range. Through the low part of this range of pin counts, high pin count pattern generators and logic analyzers become the primary tools for characterization. As the pin counts continue to grow, integration of these components in a single box occurs in the form of the application specific integrated circuit (ASIC) verification tester.

For high pin count devices (greater than 300 pins), there remain cases such as continuity and simple parametric failures, where the automated curve tracer can still be useful. However, where control of a significant number of pins is required, the use of ATE or production test equipment is increasingly indicated [11]. The issues of correlation and set-up time for ASIC verification testers make them significantly less effective in this pin count range. In general, the greater the pin count of products, the less effective the lab scale testers become versus production testers. For ultra high pin count devices with more than a thousand pins, ATE is commonly used for all electrical functions as a cost-effective method for all electrical measurement. For FA of a single high-volume product, the usefulness of the ASIC verification tester may extend well beyond this pin count range. For diverse product mixes, they tend to become limited at lower pin counts.

An additional consideration is that the electrical stimulus must be compatible with the physical failure site isolation tools. The device must be biased with the failure site isolation tool. For static DC biases, this can be achieved in a number of ways including cable harness fixtures even for relatively high pin count devices. For functional failures, this is a much more complex issue since cable can add significant performance complexity to correlation issues. While the impact of cable harnesses on the analysis of timing failures is fairly clear, the timing of edges can also be critical to assure achieving the expected electrical state during electrical preconditioning.

29.2.3 Die Exposure Techniques

In addition to a suitable electrical stimulus, failure site isolation typically requires some level of sample preparation in order to expose the die [12]. Analysis of wafers or unpackaged dice is the obvious exception. Historically, this has been decapsulation or delidding. Decapsulation is the process of

removing mold compound from the die surface of a plastic encapsulated device. Typically, the mold compound is removed only in the area above the die and bond wires. Hot fuming sulfuric and nitric acids are the most commonly used decapsulating agents. Jet etch systems [13] employing these acids or mixtures of these acids have become the dominant method for decapsulation. Jet etchers provide a safer alternative to the older techniques described below as well as a more consistent delivery of the hot decapsulation agent. A low-level vacuum is used to create the jet and hold the inverted device in place. The hole in the fluorocarbon vacuum seal for the device provides a masking effect to control the size of the decapsulation hole. The consistent delivery of hot acid is essential for reducing the decapsulation time. In addition to impacting productivity, long exposure of the mold compound to the acid can result in absorption of the acid. This can lead to swelling of the mold compound and mechanical damage to the bond wires. Older techniques, which employed immersion in a decapsulating acid or dropping the acid onto a cavity prepared in the device may also be used [14]. In general, the goal of decapsulation is to expose the top surface of the die while maintaining electrical continuity (Table 29.3).

In cases where it is more important to maintain the chemical integrity of the internal surfaces for analysis, dry decapsulation may be employed. This is particularly an important approach in cases where metallization corrosion is observed or suspected. Thermomechanical decapsulation [15,16] is a specialized procedure typically reserved for failures, which are expected to be due to metallization corrosion. While corrosion analysis has been the primary thrust of thermomechanical decapsulation, use on failures prone to recovery with wet chemical decapsulation has also occurred. The primary disadvantage of this approach is the loss of electrical continuity. The primary advantage for corrosion failures is that contamination or corrosion products are not removed from the die surface by dissolution, allowing chemical analysis of the corrosion initiators and byproducts. In the case of decapsulation recoverable failures, this approach may yield a higher percentage of devices, which continue to failure site isolation electrically. A variety of approaches have been employed which involve various combinations of grinding, breaking, and heating. One technique employed is to crack a heated device along the upper lead frame to mold compound interface. If the die surface is not exposed by the fracture, the top of the device is heated until the mold compound softens and the die can be lifted out with tweezers. In an earlier version, the backside was ground away to expose the die. The device was heated and the die lifted or pried out of the mold compound. A third approach has been to heat the device in a fume hood until it begins to smoke and twist the package. Other variations have been employed with varied success rates. Techniques are often selected based on the sample size and required success rate.

Plasma decapsulation has also been used for plastic packages [17]. It has the potential advantage of high selectivity typically using an ashing plasma (primarily oxygen with some fluorocarbon). This allows minimized etching of the die and lead frame materials. The primary disadvantage is the time required. Since filler material in the mold compound is typically not etched by the oxygen plasma, the etching of the polymer is slowed, as filler material is exposed. This makes it essential to remove the exposed filler

TABLE 29.3 Summary of Standard Approaches for Exposing the Die in Various Package Types

Package Type	Method	Comments
Plastic encapsulated IC	Jet etch	Required for many modern mold compounds Flexibility of various acids and mixtures of acids
	Dry decapsulation	Thermal mechanical fracture of package to expose die without chemical—useful for chemical analysis follow-up
	Plasma decapsulation	Time consuming approach used for niche applications where acid is altering the failure mode or defect
	Laser decapsulation	Expensive and difficult to control damage to the device being decapsulated
Cavity package	Mechanical	Both grind off lids and lid seal fracture knife or razor edge
Backside	Mechanical	Combination grinding and mechanical followed by polish to a mirror finish

material periodically. Recent efforts with laser decapsulation have also been attempted. Efforts have been limited by the cost of equipment and ability to completely remove the mold compound and maintain electrical performance of the device [18,19].

Delidding is employed for cavity packages [20]. There are a variety of techniques available. The delidding of hermetic parts is normally a relatively straightforward mechanical process of either breaking the seal or grinding off the lid or lid seal. For ceramic lid devices, the choice between breaking or grinding is usually determined by the extra risk to maintain lead integrity in breaking the lid seal versus the longer time required to remove a lid by grinding. Fixtures with knife-edges for cracking open the lid seal on many packages provide a relative safe and convenient approach. Metal lid seals can usually be fractured by tapping a sharp edge such as a razor blade into the seal. Mechanical milling may be required if access to the edges of the lid is limited e.g., cavity down pin grid arrays.

As flip-chip mounting has become more important, backside failure site isolation techniques are more commonly used. Backside techniques are also increasing in popularity for the analysis of wire-bonded devices as the number of metallization levels increases on the die. The increase in the number of metallization levels greatly reduces observability of the die. Backside analysis requires that the back of the die be exposed and usually polished in preparation for backside analysis. These techniques range from the very simple to the complex procedures. In many cases, it is possible to use standard cross-sectioning equipment to polish the back of the die. Where localized polish is required for thinner silicon, tools range from mechanical mills and ultrasonic mills to focused ion beam (FIB) and laser enhanced silicon removal processes can be employed [21–23]. In addition, the desirability of backside analysis for devices with many levels of metal or high metal area coverage, techniques have been developed to repackage wire bonded devices for backside analysis. Typically, the device is thinned to a thickness of approximately 100 μm for backside analysis. Since many of the techniques for backside analysis employ an infrared (IR) light probe, it is desirable to reduce the absorption of IR light in the sample, improving signal to noise. On the other hand, thinning to 100 μm generates devices, which retain reasonable mechanical stability. In cases where further thinning is required, FIB and chemically enhanced laser removal are most commonly used for local thinning of the device.

29.2.4 Global Techniques

Global Techniques provide a methodology for quick isolation of failure location. The two most dominant techniques have been hot spot detection (thermal) and emission microscopy [24,25]. In addition, a number of techniques also have been developed to take advantage of particular types of defects, which can be impacted by localized heating or carrier generation.

29.2.4.1 Hot Spot Detection

Historically, the earliest forms of hot spot detection were spatially resolved IR mapping and a boiling freon technique. These generally lacked good spatial and thermal resolution. With the advent of double level metallization, a global thermal approach became critical for detection of interlevel metallization shorts and liquid crystal was developed as an excellent approach for detection of these resistive shorts as well as defects at the transistor level. The thermally sensitive liquid crystal is applied to the surface of the failed device, typically using a volatile solvent to spread the liquid crystal more uniformly. Typically a liquid crystal with a thermal transition just above room temperature is ideal for hot spot detection [26–29]. A liquid crystal commonly known as K-18 has become the standard liquid crystal employed, although liquid crystals with other transition temperatures are used for failures at specific temperatures for higher power devices. It has a transition at approximately 30°C. Spots in the order of few microns are achievable with power sensitivity in the order of 100's of microwatts. The hot spots appear as a dark region in the liquid crystal. Polarizing lens enhance the visibility of the spot. Applying bias in slow AC fashion allows some control of the spot size with duty cycle as well as enhancing visibility by creating a "blinking" effect. Thermal sensitivity can be enhanced by temperature control. Ideally, the liquid crystal should be controlled at a temperature just below the transition point. This reduces the power required to

create an observable transition, effectively increasing sensitivity. Liquid crystal proved to be a powerful technique for typical leakages on 5 V devices. As power dissipation that can be generated at a defect site has fallen with decreasing power supply voltages, thermal techniques have become somewhat less useful, although they remain a critical part of the suite of global failure site isolation tools. In addition, reduction in feature size has made the large spot observed inadequate for good fail site isolation.

Fluorescent Thermomicrographic Imaging (FMI) has been used for hot spot detection [30,31]. A material, primarily EuTTA, with a high thermal coefficient of fluorescence is applied to the device, much as liquid crystal. Fluorescence maps are generated using charge coupled device (CCD) cameras. The fluorescence image of the biased failing devices is mathematically compared to the fluorescence image from the unbiased die on a pixel by pixel basis. Hot spots are indicated by areas of reduced fluorescent intensity.

Black body radiation measurements have also been used for many years for measurement of localized temperature. This was originally somewhat limited by spatial resolution and used primarily for localizing heat on printed circuit boards. The spatial resolution and thermal sensitivity of this approach have been recently improved to make this approach more applicable to die level failure site isolation. The ability to use such techniques from the backside to observe thermal behavior in the active areas of devices has also contributed to greater use of this thermal measurement approach [32].

29.2.4.2 Emission Microscopy

Emission microscopy was also developed [33] to address failure site isolation issues with double level metallization. In general, light emitted due to electron-hole recombination and hot electrons is detected by a light amplifier, which maps the location of the light intensity [25,34,35]. The resulting emission image is overlaid onto an optical microscope image. This precisely identifies the location of the emission and in most cases the defect site. It should be noted that defects away from an emission site could be the cause of emission. For example shorted metallization can result in saturated transistors, which will emit at the saturated transistor but not at the site of the short. Emission microscopy has seen limitations due to the rapid increase in the number of metallization layers. Since metal is opaque, the increase in layers significantly reduces the observability of emission sites. Emission sites, however, are observable from the backside of a device. This makes emission microscopy a key tool for backside FA. Several factors are important in utilizing emission microscopes from the backside [35,36]. The first is transmission of IR light through the silicon. For P type material, the absorption of IR light increases rapidly with doping level. Thus, higher substrate doping levels increase the importance of thinning. Another factor is an important consideration for emission microscopy from both the front and back. In general, stronger emissions have been reported in the IR region than in the visible region for many of the common emission mechanisms.

29.2.4.3 Charging Based Global Techniques

Two techniques to observe open circuits have been developed which are based on charging the interconnects in CMOS circuits with an electron beam. Charge induced voltage alteration (CIVA) [37] and Low energy charge induced voltage alteration (LECIVA) [38] have been used to identify open circuits. The technique is predicated on the assumption that CMOS signal traces predominantly connect a source/drain contact to a gate of another transistor. Generally, any charging on such a trace will be bled off at the source/drain contact. However, if the line is open, the gate side will charge, altering the behavior of the transistor. A key feature of these techniques is operation of the device at a constant current, monitoring voltage changes. This facilitates measurement and makes the techniques more sensitive. Light induced voltage alteration (LIVA) [39] is a laser based equivalent to CIVA.

29.2.4.4 Thermal Generation Based Global Techniques

Several techniques based on heating interconnects with a scanning laser beam have been developed. Optical Beam Induce Resistance Change (OBIRCH) [40,41] is a technique for detection of resistive elements. It detects resistance changes in the metallization due to localized heating. In areas without anomalies, thermal redistribution is uniform and effectively reduces the temperature. In an area with a

discontinuity, heating occurs with a greater change in resistance than in the continuous area. It is assumed that defects, which create an open or high resistance in a line, also reduce the thermal conductivity in the area. Since this is a laser-based technique, it is possible to perform OBIRCH from the backside using an IR laser. In any event, IR lasers (wavelengths with energies below the bandgap of silicon) are required for these thermal analysis techniques in order to prevent photocurrent generation. Seebeck Effect Imaging (SEI) has also been reported using the Seebeck effect to generate potential on the open portion of a net connected to a gate. The effect is monitored using constant current biasing and measuring the voltage changes similar to the xIVA approaches discussed above. Thermally Induced Voltage Alteration [42] (TIVA) has been used in a similar fashion to detect power consumption changes due to shorts in metallization. Thermally induced voltage alteration is very similar to OBIRCH, but using the constant current methodology to enhance sensitivity.

Additional techniques using heating lasers have developed to observe different electrical characteristics of the device. These include techniques which map electrical characteristics of a device, which may not be simple measurement. A common example is measurement of a pass–fail condition, which has been shown to be an effective method for resistive via isolation [43–45]. Many acronyms have been generated using somewhat similar approaches. The common feature is the monitoring of some electrical characteristic, as the heating laser is rastered over the device under test.

29.2.4.5 Carrier Generation Based Techniques (Semi-Global)

The generation of carriers in silicon can be used to modify the behavior of the circuit while in operation. When the carriers are generated well away from a junction, recombination of the carriers is likely with no net impact on the operation of the devices. However if the carriers are generated near a junction, the carriers can result in a net increase in current flow, which can be monitored as a power supply current. Two sources of carrier generation have been employed. Electron beam induced current [46] (EBIC) uses an electron beam while optical beam induced current (OBIC) [47] uses a light beam, typically a laser. These techniques have not been broadly used on very large scale integration (VLSI) devices for a number of reasons. They are not classical global techniques since an understanding of the circuit of operation is required to interpret the results as well as to determine the sites to irradiate. Electron beam induced current requires that the electron beam penetrate to the silicon in order to generate carriers. This also becomes impractical as the number of layers of metallization increases. OBIC suffers from similar disadvantages on VLSI devices because the metallization is opaque to light. However, OBIC may become more popular as a technique for backside analysis since access to the active silicon areas is not blocked by opaque metallization from the back. Light near the bandgap of silicon will transmit through silicon and have enough energy for carrier generation. This will make OBIC potentially a useful backside analysis technique.

Progress in magnetic imaging has led to the possibility of tracking currents on the device with adequate spatial resolution to accurately identify the current path on a layout. Most of the activity in this area has been on package level shorts (see below) and leakage, but is finding increased applications at the die level [48–50].

29.2.5 Probing

Probing is the oldest method for failure site isolation [51]. Generically, probing is the measurement of the electrical signals within an IC under test. This can range from contact measurement with a mechanical probe or noncontact measurement with an electron beam probe or optical methods.

29.2.5.1 Mechanical Probing

Mechanical probing has remained as a valuable tool in the FA of functional failures. Computer aided probe placement has served to extend the life of the optical microscope-based mechanical probe into the sub-micron regime although probing is difficult. The rapid increase in typical device complexity has made mechanical probing, a less desirable approach for isolating VLSI failure sites due to the increase in

the number of signals that must typically be probed. The simultaneous decrease in feature size has served to increase the difficulty of contacting a particular signal trace as well as to increase the impact of loading on measurements. High impedance probes can overcome some of the loading issues. Typically, field effect transistor (FETs) are used at the probe tip to provide the high effective impedance. Increase in device complexity have been overcome through the use of computer aided design (CAD) navigation (using the CAD database or lay-out with automated stage movement to "drive" the device to a predetermined location or to precisely locate an anomalous event such as light emission) with higher resolution stages to assist in locating the nodes that must be probed. Once the defect has been localized, it is essential to be able to track that location through subsequent analysis processes. Relocating the identified defective areas in VLSI devices in other failure tools is a time consuming task without CAD navigation. Contacting difficulties have been in part alleviated with computerized probe control and migration from optical microscopes to scanning electron microscope (SEM) and FIB based probe systems in order to improve resolution and depth of field issues. In addition, atomic force microscope (AFM) based methods have been developed. The radius of probe tips has become smaller to allow contacting the finer pitch metallization. For all of the advancements, mechanical probing of a large number of nets remains a tedious and difficult task. In spite of the problems, mechanical probing remains a very important part of the failure analyst's arsenal of tools. It is particularly useful in isolating problems, which require precise measurement of DC voltage. This type of problem arises regularly in the analysis of analog devices, while under power and in the characterization of failed circuit components. In addition, glitches and single shot events are not readily detected by electron or optical beam probing since they function primarily as sampling oscilloscopes. It is also an invaluable tool in the characterization of failed components of an integrated circuit. With current technologies, an increase in the number of failures, which are the result of transistor shifts or transistor process margin has led to an increased interest in the probing of transistors after partial deprocessing [52,53]. An alternative for topside probing is the scanning probe microscope for electrical measurements. Localized measurements of electrical properties such as capacitance and spreading resistance are possible with AFM derivatives. Comparisons of doping profiles and detection of leaky and open contacts are among the many features which can be detected [52–54].

29.2.5.2 Electron Beam Probing

Electron beam probing provides an essentially load-less probe with high bandwidth. E-beam probe capability covers a broad range of applications. In its simplest form, it is the qualitative observation of high and low voltages as reflected in dark and bright conductors respectively. Most e-beam effects are best observed at low electron beam energies (typically about 1 KV). In this range, the surface remains nearly electrically neutral i.e., incident electrons absorbed at the surface are offset by secondary and back-scattered electrons coming from the surface. By the addition of AC device operation, dynamic voltage contrast can provide information about the AC performance at various nodes. By adjusting the SEM sweep rate and frequency of device operation, a striped appearance of the AC biased portions of the circuit is achieved. Hence, it is possible to determine qualitatively where the device is toggling and where it is not operating at the appropriate frequency.

Other enhancements to basic voltage contrast effect include image processing capability and quantitative measurements of voltages and waveforms. E-beam probe systems are used routinely as noncontact sampling oscilloscopes. By operating as a sampling oscilloscope, the impact of surface charging may be further offset. As with any sampling oscilloscope, a triggering input is required with a repetitive pattern set loop.

Image processing enhancements for the most part use comparisons of good and bad devices, operating in the same mode. Image subtraction results in identification of areas where different voltage levels occur on the good and bad device. Operation of systems from workstations also allows direct access to design databases and use of the database in navigation around the circuit. Modern systems employ CAD navigation to quickly locate nodes and measure waveforms or extract timing information for all of the techniques discussed. Measurement is similar to a digitizing oscilloscope. This means that all

of the triggering and looping issues associated with a digitizing oscilloscope are factors in electron beam probing.

Probe point extraction has become an issue as the number of levels of metallization has increased. Historically, electron beam signals can be measured through one and very occasionally more layers of dielectric by capacitive coupling. Probing through passivation or probing the top metal layer minus one on depassivated devices had been routinely possible. A second useful approach to probing underlying metallization has been to anisotropically remove dielectric covering all the layers of metallization. This approach is also limited by overlying metallization spacing for both mechanical and electron beam probing. Mechanical probing is limited by spatial restriction in getting a probe tip through the upper metallization. Electron beam probing is limited by cross talk from the upper metallization. In addition, the large-scale removal of dielectric has a significant impact on the parasitic capacitance within the IC and hence device operation, particularly speed. The capabilities of a FIB are ideally suited to bringing points to the surface for either mechanical or electron beam probing. Focused ion beam can generate high aspect ratio holes to the underlying metallization. The metal deposition feature of the FIB can be used to fill the hole and create small pad for mechanical or e-beam probing. The FIB can also be used to mill holes from the backside on flip-chip devices. These processes are often facilitated by the more rapid chemically enhanced laser removal.

There are several limitations for probing on leading edge technologies. The first is the increase in the number of metallization layers. This will reduce the accessibility of nodes even with FIB point extraction. The second is the increased use of flip-chip attachment. This will make at-speed probing of devices impractical except from the backside due to difficulties in providing an electrical stimulus. A third factor, which will particularly impact electron beam probing, is cross talk [51].

29.2.5.3 Isolating the Component

Mechanical probing also remains a powerful tool for characterization of isolated failed components. Electrical characterization of failed components is often critical to understand the root cause of a failure. This may be in the form of transistor leakage measurement, measurement of current drive of a transistor, verification of a high via resistance, or verification of a metal open or short. True isolation of a failed component may require cutting conductors, isolating the failed component from the remainder of the integrated circuit. It may also entail making probe contacts to conductors on different thin film layers. A wide variety of tools have been developed for severing conductors in order to allow isolation of failure sites. The classical method for severing conductors is laser ablation. Its primary advantages are relatively low cost, speed, ease of use, and reasonable control of damage. Severing broad stripes may be difficult due to heat sinking by the metal but nibbling at the broad stripe can be effective. Spatial resolution into the 0.5 μm regime has been demonstrated but is clearly preferred for most deep sub-micron work. Ablative laser systems provide an additional capability for removing selected areas of dielectric in order to provide opening to contact metallization. In cases where failed components recover during deprocessing, parallel polishing may also be employed to expose the conductors for mechanical probing. In fact, for deep sub-micron technology, the preferred methodology is to polish to a contact or via level to expose contacts or vias for direct contact. Slight etching of the dielectric is commonly used to make the contact or via rise slightly above the flat surface. The FIB provides a wealth of capability for such probing. It cuts conductors allowing component isolation. It provides selected area deprocessing of both dielectrics (to expose conductors) and metallization (to perform work through buses). It creates probe points by drilling high aspect ratio holes to a conductor, filling the hole with metal and expanding the top to form a probe point [22,55,56] from either the topside or backside. In FIB, a gallium ion beam is rastered across a sample resulting in sputtering the surface away. The diverse applications of FIB have been largely arisen with the use of gases, which are bled onto the sample during the beam rastering. The earliest gases were organometallic compounds of tungsten and platinum that the gallium beam decomposed to form conductors. When combined with the natural cutting capability of FIB, this allowed for the rewiring of circuits. This has become a critical part of design debug. It allows repair of identified design problems.

TABLE 29.4 The Various Gas Enhanced FIB Processes Are Summarized

Gases	Application	Typical Uses
Organometallics	Metal deposition	Metal interconnects for rewiring Sacrificial layer to avoid rounding in section near the top surface
Tetraethylorthosilicate (TEOS)	Oxide deposition	Dielectric to allow more complex rewiring such as passivation removal of areas of buses
Halogens	Enhance dielectric etch	High aspect vias for rewiring
Xenon difluoride	Enhanced metal etch	Metal removal
Water	Enhanced organic etch	Sections of photoresist Selected area removal of Polyimides
Proprietary etch processes	Copper and low-*k* dielectrics	Advanced material removal

This allows the debug to continue on the repaired devices without the processing of additional silicon through a wafer fab. This drives high confidence in redesign success. It has also allowed customers to obtain prototype samples much more quickly. Recently, the addition of oxide deposition capability has greatly expanded the potential for rewiring. Enhanced etching (see Table 29.4) has also improved the capabilities for device repair. In enhanced etching, gases are bled onto the sample, which react with the surface to form gaseous compounds. The gaseous compounds can be removed much more rapidly than simple sputtering allows and etching is enhanced. These gases also provide a level of selectivity since they may react with one material to generate a gas and not with another. The FIB is also a high-resolution cross-section tool as described below.

29.2.5.4 Backside Probing

As flip-chip becomes more popular, backside acquisition of signals has become a requirement. The first approach developed was to use a FIB to mill a backside hole into field oxide areas and use e-beam probing of the exposed node (alternatively mechanical probing of a FIB constructed probe pad). This method suffers significant limitation in terms of time required to create the FIB holes and the number of nets, which can be accessed without damaging transistors. Measurement of waveforms (Laser Voltage Probe or LVP) using an optical probe, which is sensitive to changes in refractive index with electrical potential based on the Franz–Keldysh effect has been developed [51,57]. Laser voltage probe has a number of advantages. Since measurements are made at the transistor rather than in the metallization network, all AC-active nodes are equally accessible (see Table 29.5 for a comparison of optical beam probing to existing techniques). The key limitation for this approach is that the incident photon beam must be focused to as small a spot as possible. Diffraction limits this to a spot, which is becoming significantly larger than a transistor. Since the shifts in diffraction properties of the silicon are small, signal to noise is a major concern and improvements in signal processing have improved the efficiency of this approach.

Another approach is an emission microscopy based technique [58], picosecond imaging circuit analysis (PICA), which detects the light emissions that occur during the changing of state of the CMOS transistors. This is a time-resolved emission microscopy approach, using the faint but normal emissions that occur due to hot carriers during state-to-state transistors in CMOS devices. Improvements in efficiency for localized photon detectors have led to the development of single point techniques, which allow timing measurements at single point with high efficiency. The ability to improve optical detection efficiency is much higher for the single spot approach. This approach has been integrated into systems which can measure timing on devices [59,60]. It is significant to note that the other techniques acquire true waveforms, the single point PICA approach provides timing information i.e., an emission peak is observed during the time the transistors are transitioning between states. Thus the timing of the transition is known but the shape of the rising and falling edges is not clear. The spatial resolution is limited by the ability to exclude light from adjacent circuitry.

TABLE 29.5 Probing Technologies Are Compared Relative to Key Issues in Probing

	Mechanical Probe	Electron Beam Probe	Optical Beam Probing	Single Point Picosecond Imaging Circuit Analysis
Contact	Mechanical—requires direct contact to conductor	Non-contact, capacitive coupling allow probing of underlying conductor	Non-contact	Non-contact, no incident beam use
Loading	Capacitive and resistive load—minimized by high impedance probes	None	IR light may provide some electron-hole pair generation	None
DC accuracy	Very high and can be enhanced by kelvin probe techniques	Poor	Poor	Poor
AC accuracy	Need to be aware of loading effects	Limited by bandwidth but not loading effects	Limited by bandwidth but not loading effects	Rise and fall time are difficult to obtain
Node accessibility	Issues with probe placement and access to underlying nodes	Probe points required for extracting buried signals	All signals accessible Limited by laser spot size	All signals accessible

An additional limitation with increasing importance is the emission at voltages below 1 V The IR imaging of both LVP and single point PICA is improved using silicon immersion lenses (Table 29.5).

29.3 Physical Analysis Tools

Physical analysis tools are those tools, which are used to identify the physical location and physical characteristics of the defect, which is responsible for the electrical failure observation. These tools include a broad range of sample preparation and observation or microscopy techniques.

29.3.1 Package Analysis

Two very important techniques are used in the analysis of package defects: x-ray and scanning acoustic microscopy (SAM). X-ray and SAM are powerful and complementary tools for observing defects in packaging [61,62]. Packaging related defects typically manifest themselves in the form of continuity or leakage failures. For this reason, continuity and leakage failures are commonly analyzed with x-ray [63] and SAM [64]. In addition, SAM is commonly performed on all surface mount devices, which have been through a board assembly process, typically vapor phase or IR reflow conditions. X-ray provides an excellent method for observing the electrical connections within a package. Typically, the metal elements of the package can be clearly delineated. The exception is aluminum bond wires, which are difficult to distinguish due to the low atomic number of aluminum. While x-ray provides visibility into the metal interconnections, SAM provides information about the adhesion of various interfaces within the package such as the mold compound to die and mold compound to leadframe interfaces.

Two additional new tools play key roles in the analysis of failure in complex packages. Time domain reflectometry (TDR) is used in the analysis of open circuits. This technique measures the time for an electrical pulse to be reflected back to a package pin or ball [65]. Similarly, the scanning magnetic microscope shows promise for the analysis of shorted devices by allowing the mapping of current through the defect path. As pointed out above, this technique can also be applied to die level shorts and leakage. A variety of magnetic sensors have been developed [48–50].

29.3.2 Deprocessing

There are three important approaches for exposing and observing defects in IC: deprocessing, parallel polishing, and cross sectioning. Deprocessing [66–70] is the chemical removal of the various thin films formed in the wafer fabrication process in order to expose the wafer fabrication defect of interest. The order of layer removal is the reverse of order of the application during fabrication. The chemical removal of the materials can be divided into three categories: metallization, dielectrics, and silicon. The processes used for deprocessing are typically rather similar to those used in the wafer fab for the removal of the same thin film. While chemically, the processes are similar, there are marked differences. Typical wafer fab processes can use a significant overetch because their endpoint is determined by selectivity. For example, a metallization etch will be stopped at a dielectric under the metallization, as no underlying metallization is left exposed. In the comparable FA deprocessing step, the metallization etch must be selective to underlying dielectric as above but must also be stopped in the vias.

Metallization removal has historically been performed using wet chemicals where possible. Wet chemical metallization etches are generally highly selective with respect to dielectrics. The use of wet chemicals is dependent on some blocking mechanism to prevent etching underlying metallization layers through vias. This is the case where barrier-adhesion layers are used as part of the metallization structure or dissimilar metals are used in the vias such as tungsten plugs. In cases where etch of underlying metallization is not blocked, either parallel polishing or plasma based metallization etching must be used. Plasma based chemistries for aluminum metallization are normally chlorine based, which add significant to the costs and safety concerns.

For dielectrics, the plasma-based removal of the materials has been the method of choice for many years. In general, etch endpoints must be time based rather than selectivity based. This occurs because the deprocessing etch must be stopped on both metallization and the underlying dielectric as well. Early plasma applications were both barrel and planar type plasma etchers. As metallization traces became narrower and undercutting became an issue, anisotropy became a requirement as well. Reactive ion etching (RIE) of dielectrics provides this greatly enhanced anisotropy. In recent years, magnetically enhanced RIE and inductively couple plasma (ICP) process have also become popular to improve control of the etch process and reduce the chances of damage to the device being deprocessed.

29.3.3 Parallel Polishing

With the poorly defined endpoints of deprocessing, particularly dielectric deprocessing, and the increased number of thin film layers to remove, it has become significantly more difficult and time consuming to fully deprocess devices. These factors along with the difficulty of avoiding artifacts with deprocessing have led to an increase in the use of parallel polishing. Parallel polishing is the process of mechanical removal of the thin films in order to expose the wafer fabrication defect, which has resulted in failure. Frequently, parallel polishing is used to do the preliminary layer removal when the defect is expected to occur at a specific level followed by deprocessing. For example, emission microscopy may indicate a transistor level defect so that all of the metallization and interlevel dielectrics can be removed prior to starting deprocessing. Maintaining parallelism across a large die is difficult. This makes good failure site isolation critical for parallel polishing, since maintaining parallelism in a small isolated area is much easier than for large areas.

29.3.4 Cross-Section Analysis

A powerful alternative to deprocessing is cross-section analysis [71]. It is particularly useful when the position of the defect is precisely known. Cross-sections tend to be more definitive about the wafer fab process step during which the defect occurred. Deprocessing can tend to distort the location and composition of defects by creating defect replicas in underlying materials. As has been observed with many of the tools and techniques of FA discussed above, the steady reduction in feature size has led to

more stringent demands on spatial resolution for defect inspection and identification. This has clearly been the case with cross-section preparation. Early cross-section preparation was by encapsulation in clear plastic and sawing, grinding and polishing on standard metallurgical tools. In fact, this process continues to be used for assembly process defects such as package cracking, bond intermetallics, and die attach integrity etc. Cleaving and fracture along crystallographic planes, also remains a useful tool. As a tool for preparing cross-sections of a general area such as memory arrays or features in the order of a few microns, it remains a quick and efficient sample preparation technique. Precision in the order of a micron has been achieved with automated cleavers. These tools include a capability for liquid nitrogen cleaving, which reduces smearing of metallization and other ductile materials.

Improved precision has been obtained with unencapsulated sectioning techniques. In this technique, the sample is attached with wax to stub, which in turn is attached to block that is supported on a Teflon bar. This assembly is mounted on a polishing wheel, where the sample is ground away or polished. The original use of this approach employed a lime glass wheel as an abrasive. Polishing on a variety of abrasives to improve the finish and for slower sample erosion is a part of the process as well. Features as small as a few tenths of a micron can be routinely cross sectioned by this method.

However, FIB cross sectioning has become the dominant method for cross-section sample for deep sub-micron features. The cross section is prepared by sputtering or milling away a large box near the feature of interest. The box is typically terraced with the deep end near the feature of interest. The section face is milled away with decreasing beam current to create a polished surface, which can be inspected by SEM or FIB. Dual column FIB systems, SEM and FIB in the same system, provide a great improvement in efficiency for cross sectioning. Blanket Ion Mills provide an efficient and low-cost alternative for less precise cross sections. A mechanical barrier is used to stop the ion beam in areas which are to be maintained and the exposed areas are etched away very quickly.

29.3.5 Microscopy

A broad array of physical characterization techniques and tools are employed at various phases of FA. These are generally non-destructive techniques used to observe physical features of the die or package. These include microscopy of various types to allow inspection and documentation of both deprocessed devices and cross-sectioned devices [72], also see Chapter on Characterization.

As feature sizes are reduced, the minimum size of killing defects (i.e., defects which cause electrical failures) is also reduced. The spatial resolution of physical analysis tools as well as chemical analysis tools must be adequate to address the smaller size. In general, FA must be able to deal with defects, which are in the order of 10%–20% of the minimum feature size. On-line tools more typically must be able to deal with defects in the order of 30%–50% of the minimum feature size. Smaller defects are commonly dealt with more advanced tools in a research environment.

Optical microscopy remains an integral part of the FA toolkit. Optical microscopy continues to have one significant advantage over other inspection tools in that defects can be viewed through transparent dielectrics. Other significant advantages are ease of use and straightforward interpretation. While brightfield and to some extent darkfield applications are the dominant uses of the optical microscope in FA, other modes find useful niches. For example, interference contrast is particularly useful in the documentation of crystalline defects after deprocessing to silicon and decoration. Fluorescence microscopy can also prove useful in the evaluation of contamination, for example observing the spreading of an organic contaminant. In addition, it provides the basis for FMI discussed above. Polarized light microscopy finds applications in such areas as enhancement of liquid crystal transitions.

Confocal microscopy provides a method for approaching diffraction limited resolution in optical microscopy but with minimum depth of field. Confocal microscopy approaches the physical limits of optical microscope spatial resolution through the use a pinhole aperture. Spinning disc configurations and computer addition of images at various heights allow merging of images from different focal planes into a composite with nearly diffraction-limited resolution and effectively higher depth field. Spinning disc systems provide real time imaging by using a large number of pinholes. Single pinhole confocal

microscopes are not real time due the time required to manually scan in the *z* direction while grabbing images. These capabilities are further enhanced by the use of lasers in laser scanning microscope (LSM) to improve acquisition times by increasing signal levels. In addition to provide optical microscope enhancement, the LSM provides an ideal platform for optically based global failure site isolation techniques such as OBIC, LIVA, OBIRCH, and TIVA, which were discussed above.

Infrared microscopy finds a growing range of applications dependent on the IR transparency of silicon and other semiconductors. Infrared microscopy is a vital element of backside side analysis, providing the only current technology for penetrating a polished silicon surface. Infrared microscope differs in several significant ways from an optical microscope. Since IR light is not "visible," it cannot be observed directly. A converter must be used to change the IR image to a format that can be viewed. Resolution of the IR microscope is poorer than a comparable optical microscope because the wavelength of light used is larger. The use of silicon immersion lenses can improve the spatial resolution of IR microscopy by about a factor of three. Also, colors or variations with wavelength do not add significantly to the interpretation of IR region images. The theoretical limit is in the order of 0.5–0.6 of the wavelength of light used. This is in the order of 0.3 μm for visible light and worse for IR. However, many of the optical microscope techniques such as the use of lasers and confocal microscopy remain applicable.

Ultraviolet (UV) light microscopy have been developed with much better spatial resolution. Since the wavelength of light is shorter, a better theoretical resolution is possible. Since the transparency of the common dielectrics, silicon oxide and silicon nitride, extend into the UV region, UV microscopy provides a useful extension of the optical microscope. Like IR microscope, it requires a converter since the UV image is not directly observable. Monochromatic UV light is normally used to reduce chromic aberration.

The usefulness of the optical microscope, however, continues to become limited, as the minimum feature size has passed the diffraction limit for the wavelengths of visible light. This has driven a marked trend towards more SEM utilization relative to optical microscopy. This is in large part due to the resolution and depth of field limitations of the optical microscope, which come into play with smaller geometry devices and an increased number of layers in devices. It is also a reflection of the increased ease use of the SEM and particularly field emission scanning electron microscope (FESEM), which has become dominant in the semiconductor industry.

The FESEM has become an essential tool for FA. The very high magnification capability is essential for documentation of defects in deep sub-micron processes. In addition, SEM provides an excellent depth of field, for low magnification inspections such as bond wires. In addition to its value as a microscope, the FESEM provides the platform for other e-beam techniques such as backscattered electron imaging, electron beam probing, and EBIC and x-ray microchemical analysis.

As killing defect sizes continue to shrink, observation and documentation of these defects require higher magnification and better resolution. This requirement has fueled the transitions from optical microscopes to thermionic emission SEM's to FESEM and it now drives greater use of higher resolution options. These requirements are filled in part by higher resolution FESEM's, commonly referred to as in-the-lens systems in which the sample is located within the electron lenses. Unfortunately, this arrangement severely limits sample size. Transmission electron microscope (TEM) and AFM also provide tools which fill portions of the resolution gap, particularly the TEM whose use has grown dramatically in recent years. The derivatives of the AFM [73,74] are commonly referred as scanning probe microscopes and include a broad range of measurement capabilities including voltage [58] (electric field), current (magnetic field), and capacitance [59] as describe above. In addition, near field optical microscopy can provide a limited optical approach for microscopy which is not diffraction limited [60,75].

29.3.6 Transmission Electron Microscopy

The need for better resolution has led to the more wide spread use of TEM for FA. The TEM is however limited to some extent by sample preparation. The TEM cross-section sample preparation techniques can be grouped into three categories: standard general area approaches, specific area approaches, and FIB

approaches, which now dominate the sample preparation arena. In addition, plan view TEM sample preparation is becoming more popular. The classical approach for TEM sample preparation has been the thinning of a stack of devices glued together followed by dimpling and ion milling. This approach is used to observe the general processing features, such as thickness measurements. Several approaches for mechanically polishing a specific area of interest from both sides have been developed as well. Although these approaches can be very time consuming, specific area sections are essential for defects such as resistive via contacts. More rapid sections can be prepared using the FIB. The FIB can be used to mill cross-section views from both sides of a defect, resulting in the thinned section required for TEM viewing. The only serious limiter for FIB sample preparation is ion beam damage to the sample from gallium ion implantation. Techniques have been identified to reduce this effect, but it is not eliminated.

29.4 Chemical Characterization

Contamination continues to be a primary cause of failures in IC. This occurs primarily in the form of particles but may occur in many other forms such as spots from evaporation and co-implanted contaminants. The total levels of contamination are normally exceptionally small and concentrated in a very small volume. When selecting a method of analysis (also see Chapter on Characterization), several factors are critical: the spatial resolution of the technique, the volume of material to be analyzed, the sensitivity of the technique to the contaminant and background elements, the type of chemical information provided by the technique, and the ease of performing the analysis. Most analytical techniques can be viewed in terms of four elements: incident radiation, the physical interaction of the incident radiation and the sample, radiation flux from the sample which results from that interaction, either a new form of radiation excited by the incident radiation or an attenuation of the incident radiation, and a detector for the radiation flux from the sample.

The spatial resolution is an exceptionally important requirement for analysis in semiconductors. Very small particles must be analyzed. In a FA environment, it is important to be able to perform analysis on particles, which are 10%–20% of the minimum feature size. In general, the spatial resolution of a technique is dominated by spot size of the incident radiation and to some extent by the dispersion of the incident radiation in the sample as in the case of characteristic x-ray analysis. The requirement to accurately focus the incident radiation tends to make charged species, particularly electrons, most useful as incident radiation for high-resolution application. This is true because charged particle beams, particularly electron beams, are readily focused. However, the use of lasers and new focusing techniques for electromagnetic radiation are bringing other techniques into prominence. The depth of analysis is controlled by one of the two factors, the depth of penetration of the incident beam and the inelastic escape probability of the radiation flux from the sample as a function of sample depth. For example, the depth of electron beam penetration controls x-ray analysis depth. However, the depth of analysis for Auger, using the same incident radiation, is controlled by the escape cross-section of the Auger electron as a function of depth in the sample.

Sensitivity tends to be dominated by several factors, which can be viewed as impacting the signal to noise ratio of the technique. The first is the cross-section of the interaction of the incident radiation with the sample. It is very difficult to get high sensitivity from low cross-section interactions. Basically, high cross-section events provide a higher signal level. The second factor is the detector and how efficiently it can collect the signal. This includes factors such as geometry, which impacts the fraction of the signal, which can be detected, and detector efficiency. It also includes the noise level of detector and how it compares with the signal level.

The type of chemical information provided is typically dependent on the type of interaction between the incident radiation and the sample. There are basically two types of information obtained: atomic and molecular. Atomic (or elemental) information is about what elements are present in the volume sampled. Molecular information defines in some form, the chemical bonding between the elements and/or at least the oxidation state of the elements present. As a generality, atomic information is most readily obtained

TABLE 29.6 The Primary Qualities of Analytical Techniques Which Impact FA Are Listed for the Techniques Most Commonly Used

Technique	Spatial Resolution	Depth of Analysis	Sensitivity	Ease of Use/ Interpretation	Data Type	Applications
Energy dispersive	*Good*	Bulk: 2–5 μm	Moderate	High	Atomic	General purpose
Auger	Good	*Surface: 20–50* Å	Poor	Moderate	Atomic	Adhesion problems
Secondary ion mass spectroscopy	Moderate	Surface	*Very high*	Low	Atomic/ molecular	Doping levels
Fourier transform infrared spectroscopy	Adequate	Varies	High	Moderate	*Molecular*	Organic contamination

Primary benefit(s) of techniques are italicized.

from interaction of the incident radiation with inner shell electrons, while molecular information is best obtained from interactions, which impact interatomic interactions such as vibration or rotational frequencies (Table 29.6).

29.4.1 X-Ray Analysis (Energy Dispersive)

Energy Dispersive x-ray Spectroscopy (EDS or EDX) remains the mainstay chemical analysis tool for FA. Its primary advantage is that it is used in conjunction with a SEM, which is already available in the FA laboratory. It has an excellent spatial resolution. Further, information can be obtained quickly and easily interpreted by the analyst. The spatial resolution is good since the SEM focuses the electron beam. The capability to move the electron beam also provides the ability for dot mapping and line scanning. The depth of beam penetration is typically several microns. Since the escape cross-section for x-rays is high, analysis is performed of material down to that depth in the sample. This makes the surface a relatively small part of the sampled volume and results in x-ray analysis being of limited value for very thin surface films. The sensitivity of the technique is in the order of 0.1%–1% for most elements but less for light elements. The primary disadvantages of the technique are limited sensitivity, especially for light elements and limited applicability to surface analysis. These do not however keep it from being the primary analytical technique for FA.

Wavelength Dispersive (WD) Analysis detects the same interaction as Energy Dispersive Analysis. Hence, the type of data collected is identical, however, the different method of detection improves the resolution of x-ray peaks and yields improved signal to noise ratios. This results in improvement in sensitivity. The drawback is that the time for analysis is long relative to energy dispersive analysis. Microcalorimeters are likely to supplant EDS detectors, providing energy resolution on a par with WD with many of the good characteristics of the EDS [76].

29.4.2 Auger

Auger analysis is also an electron beam technique. The observed radiation is Auger electrons, which result from a two-electron process. Since the inelastic escape cross-section from below the surface is low, Auger electrons are detected only from the top 10 to 30 Å. Layers below the surface can be analyzed by sputter etching away the surface layers, typically in 50–500 Å increments. Depth profiling is possible by analysis of sequential spectra. In addition, some limited molecular information can be obtained from shifts in the energy of the Auger electrons with oxidation state. The primary limitation of Auger is low sensitivity. For surface analysis, it is a significant improvement over x-ray analysis, since the volume of material analyzed is the surface only. Auger application is best in detection and

thickness estimation of thin film contamination. Thus Auger is an ideal technique to study delaminations and disadhesions. It is commonly used to detect contamination on a bond pad. Great caution is required in interpretation of data, particularly initial data which frequently only identifies adsorbed material on the surface. The use of controls, known good material, should be standard, a part of all, but the most routine Auger analyses.

29.4.3 Secondary Ion Mass Spectroscopy

Secondary ion mass spectroscopy (SIMS) is a very powerful FA tool when a requirement for high sensitivity occurs. The technique uses an incident high-energy ion beam. This beam sputters away surface atoms and atomic clusters, some of which are emitted as ions. These ions, both negative and positive, can be analyzed via a mass spectrometer. For depth profiling, the ion beam is rastered over the area of interest in a pattern of reducing size to minimize edge effects. The spatial resolution of SIMS can be somewhat limited by the requirement to begin with a large area raster. Some molecular information is available from the atomic clusters which are sputtered. Secondary ion mass spectroscopy is the most sensitive (in the order of 20 ppb) of the routinely used FA techniques. This makes the use of control samples essential, since many elements will be detected at very low levels even on the cleanest of devices, making interpretation of data difficult. While SIMS is essentially a surface technique, it continuously sputters away the surface. This makes depth profiling, a natural application of SIMS and in fact, this is the output format normally obtained with only mass peaks of interest plotted. In addition, laser based techniques which analyze ions sputtered from a surface have also been developed. The techniques summarized above are the workhorse tools for FA. They have the common advantage of utilizing a charged beam, which results in good spatial resolution. Unfortunately, they share the common limitation of providing little or no molecular information. The techniques summarized below fill this gap in some manner and can be employed in FA as required.

29.4.4 Microspot Fourier Transform Infrared Spectroscopy

Fourier Transform Infrared Spectroscopy (FTIR) measures absorption of IR light by a sample. Absorption of IR light is measured with an interferometer with the spectrum expressed as the Fourier transform of interferogram. Spectra can be obtained from solvent extraction of a sample as well as directly from the sample. The development of Microspot techniques has greatly enhanced the applicability of this technique to FA. Now areas of a few microns across the organic contamination can be successfully analyzed in a reflection mode. The technique is most useful in FA for the identification of organic contamination. Organic compounds have rotational and vibrational absorptions in the IR region. In most cases, the contaminant FTIR spectrum is compared to libraries of organic compounds or spectra of suspected contamination sources. Micro-Raman spectroscopy has also been used and provides a good complement to the FTIR technique.

29.4.5 Others

A number of other analytical techniques may be useful from time to time in FA. Some of these fill specific niches in analytical capability. The x-ray photoelectron spectroscopy (XPS) observes electrons, which are ejected from atoms after interaction with monochromatic x-rays. Elemental characterization is by binding energy (x-ray energy minus the kinetic energy of the electron). Oxidation state information is obtained from energy shifts with oxidation states. This is also a surface technique, since as with Auger, the probability of electron escape without energy loss from below the surface is low. The primary limitation of the technique is the limitation of focus of the incident x-ray beam. Although significant progress has been made in columnating x-ray beams, the spatial resolution remains inadequate for many FA applications. The technique may be used in conjunction with Auger analysis in order to compliment the elemental information of Auger with molecular information of XPS. Other techniques with FA

applications include Rutherford Backscattering Spectroscopy, Ion Chromatography, Thermal Gravimetric and Thermal Mechanical Testing. Specifically, Rutherford Backscattering Spectroscopy can be used to measure stoichiometry, particularly in silicides. Ion Chromatography of water extraction has been useful in measurement of total ionic surface contamination on samples such as packages and wafers with relatively large surface areas. Thermal Gravimetric and Thermal Mechanical Testing are required to evaluate the curing of mold compounds. Even this lengthy listing omits such techniques as Atomic Absorption and x-ray Diffraction Techniques. In addition, as TEM use has increased, Electron energy loss spectroscopy (EELS) has become a valuable analytical tool, complementing EDS which can be put on a TEM as well as on a SEM. Obviously, one challenge of FA is to select the analytical techniques, which are best suited to a particular situation.

29.5 Future of Failure Analysis

With the decrease in life cycles in products and technology, the emphasis in diagnostic activity will continue to be on more upstream activities (see Figure 29.3). Design debug capability to quickly assure second pass silicon success will be essential for supporting customer requirements. Short life cycles will also drive a requirement to ramp yield more quickly. The short life cycle will also reduce the acceptability of long qualification processes, which result from qualification failures. A long-term reduction in qualification failures coupled with an improved reliability should drive fewer qualification failure analyses. The customer return analysis flow is a very lagging indicator of quality and reliability performance and should become a less significant part of the improvement process, particularly for products with short lives.

In addition to moving diagnostic activities more upstream, the technology roadmaps will drive significant changes in the tool requirements for FA. The development of tools for FA is driven by both device and process complexity as well as changes in the packaging technology. Increases in the number of levels of metallization and transitions to flip-chip are driving the broader use of backside analysis tools. Smaller feature sizes are driving demands for tools with higher spatial resolution in both the fab and assembly arenas. In addition, the industry is rapidly transitioning to many new materials such as copper metallization and low-*k* dielectrics currently and high-*k* gates and porous dielectrics in the forecast. Changes in FA technology have historically occurred in several ways. Abrupt changes typically result from radical changes in the technology. For example, surface mount package suddenly drove a strong

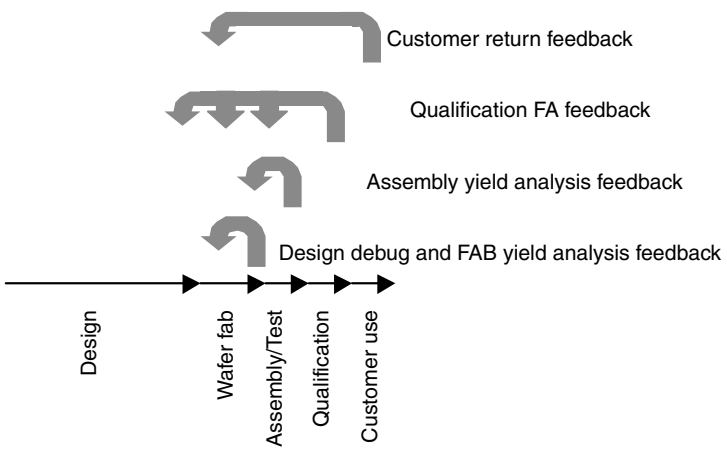


FIGURE 29.3 Failure analsis (FA) feedback loops are illustrated. Future emphasis is expected on shorter feedback loops.

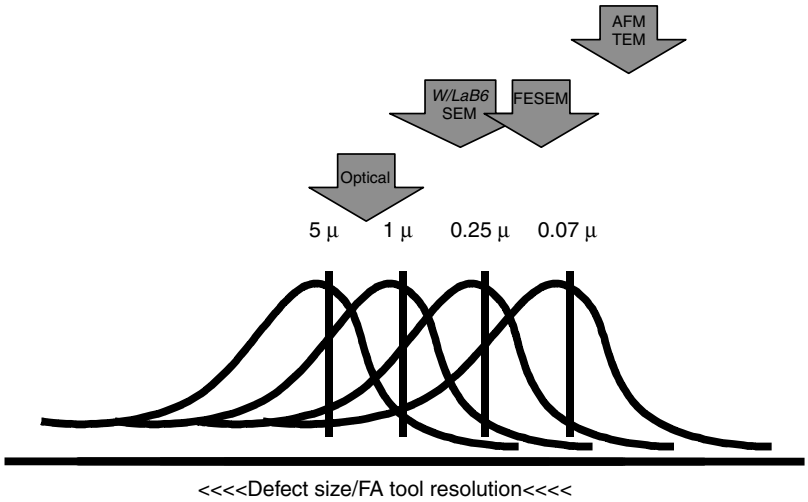


FIGURE 29.4 Arbitrary distribution of defect size by technology node versus the resolution of tools is illustrated.

requirement for acoustic microscopy. Similarly, double level metal drove much of the need for global failure site isolation. Currently, flip-chip assembly and the rapid increase in the number of metallization levels are driving backside analysis technology. On the other hand, FA tends to look at a distribution of failures, giving rise to a gradual need for improved resolution in tools. This is illustrated for inspection of defects such as particles in Figure 29.4. It is apparent that the resolution of current tools will continue to be challenged as technology feature sizes and the corresponding size distribution of killing defects continues to shrink. The same type of effect is driven in FA by increase in device complexity. This includes the migration from bench test to specialized FA testers to use of ATE.

References

1. Wagner, L. "Introduction." In *Failure Analysis of Integrated Circuits: Tools and Techniques*. 1–10. Boston, MA: Kluwer Academic Publishers, 1999.
2. Lam, D., and Y. K. Swee. "Electrical Failure Analysis in High Density DRAMs." In *IEEE International Workshop on Memory Technology, Design and Test Symposium*, 24, 1994.
3. Butler, K. M., K. Johnson, J. Platt, A. Jones, and J. Saxena. "Automated Diagnosis in Testing and Failure Analysis." *IEEE Design and Test* 14, no. 3 (1997): 83.
4. Saxena, J., K. M. Butler, H. Balachandran, D. B. Lavo, B. Chess, T. Larrabee, and F. J. Ferguson. *Proceedings IEEE International Test Conference*, 887, 1997.
5. Pore, M., G. Gilfeather, and L. Levy. "Risk Assessment in Signature Analysis." *Proceedings International Symposium for Testing and Failure Analysis*, 177, 1996.
6. Lakshminarayan, C. K., S. Pabbisetty, O. Adams, F. Pires, and M. Thomas. "Signature Analysis: Statistical Models and Their Application to FA." *Proceedings International Symposium for Testing and Failure Analysis*, 183, 1996.
7. Frank, S., W. Tan, and J. F. West. "Electrical Characterization." In *Failure Analysis of Integrated Circuits: Tools and Techniques*. 13–41. Boston, MA: Kluwer Academic Publishers, 1999.
8. Riordan, W. C., R. Miller, and E. R. St. Pierre. "Reliability Improvement and Burn in Optimization through the Use of Die Level Predictive Modeling." *Proceedings International Reliability Physics Symposium*, 431, 2005.

9. Seo, J. S., S. S. Lee, S. Daniel, and C. K. Yoon. "Temperature Dependence of Quiescent Currents as a Defect Prognosticator and Evaluation Tool." *Proceedings International Symposium for Testing and Failure Analysis*, 245, 1996.
10. Beall, J., and D. Wilson. "Curve Tracer Applications and Hints for Failure Analysis." In *Microelectronics Failure Analysis Desk Reference*. 3rd ed., 25–39. Metals Park, OH: ASM International, 1993.
11. Patrick, D. S., L. C. Wagner, and P. T. Nguyen. "ATE Failure Isolation Methodologies for Failure Analysis, Design Debug and Yield Enhancement." *Proceedings International Symposium for Testing and Failure Analysis*, 235, 1998.
12. Ngo, P. "Die Exposure." In *Failure Analysis of Integrated Circuits: Tools and Techniques*. 59–66. Boston, MA: Kluwer Academic Publishers, 1999.
13. Tang, P. "Techniques and New Etch Block Design to Enhance the Jet Etch Decapsulation." *Proceedings International Symposium for Testing and Failure Analysis*, 134, 1985.
14. Lee, T. W. "Chemical Decapsulation Revisited." *Proceedings International Symposium for Testing and Failure Analysis*, 113, 1987.
15. Wagner, L. C., S. Boddicker, P. Ngo, D. Morgan, and T. Myers. "Failure Analysis of Plastic IC Package Integrity and Related Failure Mechanisms." In *New Technology in Electronic Packaging*, 353. New York: ASM International, 1990.
16. Corum, D. L. "Mechanical Decap Method for Plastic Devices." *Proceedings International Symposium for Testing and Failure Analysis*, 95, 1984.
17. Tomasi, D. "Failure Analysis Applications of Plasma." *Proceedings International Symposium for Testing and Failure Analysis*, 35, 1987.
18. Skoglund, L., and R. Dias. "Laser Milling Techniques for Stacked Die Package Applications." *Proceedings International Symposium for Testing and Failure Analysis*, 369, 2004.
19. Carter, G. "Laser Decapsulation of Transfer Molded Plastic Packages for Failure Analysis." *Proceedings International Symposium for Testing and Failure Analysis*, 117, 2002.
20. Bruenderman, T. "Opening Techniques for IC Ceramic Packages." *Proceedings International Symposium for Testing and Failure Analysis*, 190, 1983.
21. Goruganthu, R. R., M. Bruce, J. Birdsley, V. Bruce, G. Gilfeather, R. Ring, N. Antoniou, J. Salen, and M. Thompson. "Controlled Silicon Thinning for Design Debug of C4 Packaged ICs." *Proceedings International Reliability Physics Symposium*, 327, 1999.
22. Lee, R., and N. Anoniou. "FIB Micro-Surgery on Flip-Chips from the Backside." *Proceedings International Reliability Physics Symposium*, 455, 1998.
23. Chiang, C., N. Khurana, D. T. Hurley, and K. Teasdale. "Backside Emission Microscopy for Integrated Circuits on Heavily Doped Substrate." *Proceedings International Reliability Physics Symposium*, 447, 1998.
24. Barton, D. "Global Failure Site Isolation: Thermal Techniques." In *Failure Analysis of Integrated Circuits: Tools and Techniques*, 67–86. Boston, MA: Kluwer Academic Publishers, 1999.
25. Cole, E. I., and D. L. Barton. "Failure Site Isolation: Photon Emission Microscopy, Optical/Electron Beam Techniques." In *Failure Analysis of Integrated Circuits: Tools and Techniques*, 87–112. Boston, MA: Kluwer Academic Publishers, 1999.
26. Burgess, D. "Liquid Crystal: Best Ideas from 15 Years." *Electronic Device Failure Analysis Newsletter* 1, no. 2 (1999): 7.
27. Burgess, D. "Advanced Liquid Crystal for Improved Hot Spot Detection Sensitivity." *Proceedings International Symposium for Testing and Failure Analysis*, 341, 1992.
28. Lindgren, J. C., L. Wagner, J. Martin, B. Carbajal, and R. Cox. "The Role of Failure Analysis in Problem Solving: Gate Oxide Integrity Problems." *Proceedings International Symposium for Testing and Failure Analysis*, 11, 1987.
29. Ferrier, S. "Thermal and Optical Enhancements to Liquid Crystal Hot Spot Detection Methods." *Proceedings International Symposium for Testing and Failure Analysis*, 57, 1997.
30. Tangyunyong, P., A. Y. Liang, A. W. Righter, D. L. Barton, and J. M. Soden. "Localizing Heat-Generating Defects Using Fluorescent Microthermal Imaging." *Proceedings International Symposium for Testing and Failure Analysis*, 55, 1996.

31. Glacet, J.-Y., and S. Berne. "A User Friendly System for Fluorescent Microthermal Imaging and Light Emission Microscopy." *Proceedings International Symposium for Testing and Failure Analysis*, 63, 1996.
32. Breitenstein, O., J. P. Rakotoniaina, M. H. Al Rifai, M. Gradhand, F. Altmann, and T. Riediger. "New Developments in IR Lock-In Thermograpy." *Proceedings International Symposium for Testing and Failure Analysis*, 595, 2004.
33. Khurana, N. "Pulsed Infra-Red Microscopy for Debugging Latch-Up on CMOS Products." *Proceedings International Reliability Physics Symposium*, 122, 1984.
34. Hawkins, C. F., J. M. Soden, E. I. Cole Jr., and E. S. Snyder. "The Use of Light Emission in Failure Analysis of CMOS Ics." *Proceedings International Symposium for Testing and Failure Analysis*, 55, 1990.
35. Barton, D. L., P. Tangyunyong, J. M. Soden, A. Y. Liang, F. J. Low, A. N. Xaplatin, K. Shivanandan, and G. Donohoe. "Infrared Light Emission from Semiconductor Devices." *Proceedings International Symposium for Testing and Failure Analysis*, 9, 1996.
36. Wu, N. M., K. Weaver, and J. H. Lin. "Back Side Emission Microscopy for Failure Analysis." *Proceedings International Symposium for Testing and Failure Analysis*, 393, 1996.
37. Cole, E. I. Jr., and R. E. Anderson. "Rapid Localization of IC Open Conductors Using Charge-Induced Voltage Alteration (CIVA)." *Proceedings International Reliability Physics Symposium*, 288, 1992.
38. Cole, E. I. Jr., J. M. Soden, B. A. Dodd, and C. L. Henderson. "Low Electron Beam Energy CIVA Analysis of Passivated ICs." *Proceedings International Symposium for Testing and Failure Analysis*, 23, 1994.
39. Cole, E. I. Jr., J. M. Soden, J. L. Rife, D. L. Barton, and C. L. Henderson. "Novel Failure Analysis Techniques Using Photon Probing with a Scanning Optical Microscope." *Proceedings International Reliability Physics Symposium*, 388, 1994.
40. Nikawa, K., and S. Inoue. "New Laser Beam Heating Methods Applicable to Fault Localization and Defect Detection in VLSI Devices." *Proceedings International Reliability Physics Symposium*, 346, 1996.
41. Nikawa, K., and S. Inoue. "Various Contrasts Identifiable from the Backside of a Chip by 1.3 μm Laser Beam Scanning and Current Changing Imaging." *Proceedings International Symposium for Testing and Failure Analysis*, 387, 1996.
42. Cole, E. I. Jr., P. Tangyunyong, and D. L. Barton. "Backside Localization of Open and Shorted IC Interconnections." *Proceedings International Reliability Physics Symposium*, 129, 1998.
43. Cole, E. I. Jr., P. Tangyunyong, C. F. Hawkins, M. R. Bruce, V. J. Bruce, W. L. Chong, and R. M. Ring. "Resistive Interconnect Localization." *Proceedings International Symposium for Testing and Failure Analysis*, 43, 2001.
44. Bruce, M. R., V. J. Bruce, D. H. Eppes, J. Wilcox, E. I. Cole, Jr., P. Tangyunyong, and C. F. Hawkins. "Soft Defect Localization (SDL) on ICs." *Proceedings International Symposium for Testing and Failure Analysis*, 21, 2002.
45. Falk, R. A. "Advanced LIVA/TIVA Techniques." *Proceedings International Symposium for Testing and Failure Analysis*, 59, 2001.
46. Cole, E. I. Jr., C. R. Bagnell Jr., B. G. Davies, A. M. Neacsu, W. V. Oxford, and R. H. Propst. "Advanced Scanning Electron Microscopy Methods and Applications to Integrated Circuit Failure Analysis." *Scanning Microscopy* 2 (1988): 133.
47. Wills, K. S., T. Lewis, G. Billus, and H. Hoang. "Optical Beam Induced Current Applications for Failure Analysis of VLSI Devices." *Proceedings International Symposium for Testing and Failure Analysis*, 21, 1990.
48. Schrag, B., X. Y. Liu, M. J. Carter, and G. Xiao. "Scanning Magnetoresistive Microscopy for Die-Level Sub-Micron Current Density Mapping." *Proceedings International Symposium for Testing and Failure Analysis*, 2, 2003.
49. Woods, S., N. M. Lettsome Jr., A. B. Cawthorne, L. A. Knauss, and R. H. Koch. "High Resolution Current Imaging by Direct Magnetic Field Sensing." *Proceedings International Symposium for Testing and Failure Analysis*, 6, 2003.

50. Crepel, O., P. Descamps, P. Poiner, R. Desplats, P. Perdu, and A. Firiti. "Magnetic Current Imaging Techniques: Comparative Case Studies." *Proceedings International Symposium for Testing and Failure Analysis*, 29, 2004.
51. Talbot, C. G. "Probing Technology for IC Diagnosis." In *Failure Analysis of Integrated Circuits: Tools and Techniques*, 113–43. Boston, MA: Kluwer Academic Publishers, 1999.
52. Liu, L., Y. Wang, H. Edwards, D. Sekel, and D. Corum. "Combination of SCM/SSRM Analysis and Nanoprobing Techniques for Soft Single Bit Failure Analysis." *Proceedings International Symposium for Testing and Failure Analysis*, 38, 2004.
53. Tong, T. X., and A. N. Erickson. "Current Image Atomic Force Microscopy (CI-AFM) Combined with Atomic Force Probing (AFP) for Location and Characterization of Advanced Technology Node." *Proceedings International Symposium for Testing and Failure Analysis*, 42, 2004.
54. Edwards, H. "Pn-Junction Delineation in Si Devices Using Scanning Capacitance Spectroscopy." *Proceedings International Symposium for Testing and Failure Analysis*, 529, 2000.
55. Ximen, H., and C. G. Talbot. "Halogen-Based Selective FIB Milling for IC Probe-Point Creation and Repair." *Proceedings International Symposium for Testing and Failure Analysis*, 141, 1994.
56. Ullmann, P. F., C. G. Talbot, R. A. Lee, C. Orjuela, and R. Nicholson. "A New Robust Backside Flip-Chip Probing Methodology." *Proceedings International Symposium for Testing and Failure Analysis*, 381, 1996.
57. Paniccia, M., R. M. Rao, and W. M. Yee. "Optical Probing of Flip Chip Packaged Microprocessors." *Journal of Vacuum Science* November/December (1998).
58. Kash, J. A., J. C. Tsang, and D. R. Knebel. "Non-Invasive Backside Failure Analysis of Integrated Circuits by Time-Dependent Light Emission: Picosecond Imaging Circuit Analysis." *Proceedings International Symposium for Testing and Failure Analysis*, 483, 1998.
59. Song, P., F. Stellari, J. P. Eckhardt, and T. McNamara. "Timing Analysis of a Microprocessor PLL Using High Quantum Efficiency Superconducting Single Photon Detector (SSPD)." *Proceedings International Symposium for Testing and Failure Analysis*, 197, 2004.
60. Ouimet, P., J. Goertz, O. Rinaudo, L. Long, and S. Yeung. "Analysis of 0.13 Micron CMOS Technology Using Time Resolved Light Emission." *Proceedings International Symposium for Testing and Failure Analysis*, 203, 2004.
61. Moore, T. M., and C. D. Hartfield. "Package Analysis: SAM and X-Ray." In *Failure Analysis of Integrated Circuits: Tools and Techniques*, 43–57. Boston, MA: Kluwer Academic Publishers, 1999.
62. Wagner, L. C. "IC Package Reliability Testing." In *Characterization of Integrated Circuit Packaging Materials*, 1–26. Boston, MA: Butterworth-Heinemann, 1993.
63. Coangelo, J. "Advanced Radiographic Techniques in Failure Analysis." In *Microelectronics Failure Analysis Desk Reference*. 3rd ed., 51–8. Metals Park, OH: ASM International, 1993.
64. Moore, T. M. "Inspecting IC Packages with C-Mode Acoustic Microscopy." In *Microelectronics Failure Analysis Desk Reference*. 3rd ed., 41–50. Metals Park, OH: ASM International, 1993.
65. Smolyansky, D. A. "Electronic Package Failure Analysis Using TDR." *Proceedings International Symposium for Testing and Failure Analysis*, 277, 2000.
66. Yim, D. "IC Deprocessing." In *Failure Analysis of Integrated Circuits: Tools and Techniques*, 145–57. Boston, MA: Kluwer Academic Publishers, 1999.
67. Lee, T. W. "Chemical Etch Formulation and History." In *Microelectronics Failure Analysis Desk Reference*. 3rd ed., 111–13. Metals Park, OH: ASM International, 1993.
68. Lee, T. W. "A Review of Wet Etch Formulas for Silicon Semiconductor Failure Analysis." *Proceedings International Symposium for Testing and Failure Analysis*, 319, 1996.
69. Beall, J. "Plasma Etching." In *Microelectronics Failure Analysis Desk Reference*. 3rd ed. 121–23. Metals Park, OH: ASM International, 1993.
70. Kiefer, D. S. "Reactive Ion Etch Recipes for Failure Analysis." In *Microelectronics Failure Analysis Desk Reference*. 3rd ed., 125–30. Metals Park, OH: ASM International, 1993.
71. Haddock, T., and S. Boddicker. "Cross-Section Analysis." In *Failure Analysis of Integrated Circuits: Tools and Techniques*, 159–73. Boston, MA: Kluwer Academic Publishers, 1999.
72. Wagner, L. "Inspection Techniques." In *Failure Analysis of Integrated Circuits: Tools and Techniques*, 175–93. Boston, MA: Kluwer Academic Publishers, 1999.

73. Bridges, G. E., and D. Thomson. "Non-Contact Probing of Integrated Circuits Using Electrostatic Force Sampling." *Proceedings International Symposium for Testing and Failure Analysis*, 169, 1998.
74. Hockwitz, T., A. Henning, C. Dagnlian, R. Bolam, P. Coutu, R. Cluck, and J. Slinkman. "DRAM Failure Analysis with the Force-Based Scanning Kelvin Probe." *Proceedings International Reliability Physics Symposium*, 217, 1996.
75. Cramer, R. M., L. J. Balk, R. Chin, R. Boylan, S. B. Kammer, F. J. Reineke, and M. Utlaut. "The Use of Near Field Scanning Optical Microscopy for Failure Analysis of ULSI Circuits." *Proceedings International Symposium for Testing and Failure Analysis*, 19, 1996.
76. Simmnacher, B., R. Weiland, J. Hohne, F. V. Feilitzsch, and C. Hollerith. *Proceedings of the 14th European Symposium on Reliability of Electron Devices, Failure and Physics and Analysis*, 1675, 2003.

30

Reliability Physics and Engineering

30.1 Introduction **30**-1
30.2 Accelerated Testing **30**-2
30.3 Time-to-Failure Modeling **30**-3
30.4 Time-to-Failure Statistics **30**-6
Lognormal Distribution • Weibull Distribution
30.5 Failure Rate **30**-10
30.6 Acceleration Factor **30**-10
30.7 Time-to-Failure Models for Selected ULSI Failure Mechanisms **30**-11
Electromigration • Contact Electromigration • Corrosion • Stress Migration (SM) • Cyclic Fatigue
30.8 Time-Dependent Dielectric Breakdown **30**-23
E-Model (Field-Based Model) • 1/E—Model (Current-Based Model) • Complementary Models (Inclusion of Both Field and Current into a Single Model) • Mobile-Ions/Surface Inversion • Channel Hot-Carrier Injection (HCI)
30.9 Summary and a Look into the Future **30**-30
References **30**-30

J.W. McPherson
E.T. Ogawa
Texas Instruments, Inc.

30.1 Introduction

Ultra-large-scale-integrated-circuit (ULSI) failures can usually be attributed to the failure of a given material under stress. This might be dielectric breakdown which can occur under electric-field stress, electromigration (EM)-induced metal-voiding failure under current-density stress, corrosion failure under relative-humidity stress, or a lifted bonding ball due to mechanical stress. The term "stress," as used here, is quite general and not restricted to the more common meaning: mechanical stress.

Stress will refer to any external agent which is capable of producing an irreversible change in the material properties such that the material can no longer function satisfactorily in its intended application [1–3]. In the case of dielectrics, this could be the dielectric breakdown which occurs when an electric field stress exceeds the dielectric strength of the material (e.g., for SiO_2 this is >10 MV/cm), or in the case of metals, this might be the rupture which occurs when a mechanical stress is applied which exceeds the fracture strength of the metal (e.g., for aluminum interconnects this is >600 MPa).

Even when a material is stored at a fixed level of stress less than the material's strength, the material will still degrade with time and failure is eventually expected. The observed time-to-failure (TF) will depend on the temperature and the magnitude of the applied stress, relative to the breakdown strength of the

material. The breakdown strength is usually defined as the level of stress at which the material is expected to instantaneously fail. [By instantaneous, it is meant that the TF is extremely short (∼m s) relative to the TF (∼years) when the material is stressed at 50% of this level]. To ensure that time-dependent failures are minimized during the expected lifetime of the product, a good engineering design will comprehend the distribution of material strengths that can be expected during normal processing and then keep the design-level/application-level well below these strength values. This is, however, becoming increasingly difficult for ULSI devices where feature sizes continue to be aggressively scaled according to Moore's Law. [Note: Moore's Law for Si-based chips states that the number of bits/functions on a chip tends to quadruple every 36 months.] Moore's Law drives a linear scaling reduction of 0.7× for device geometries per technology node. This scaling has led to higher current densities in the metal conductors, greater Joule heating, and higher operating electric fields in the dielectrics. How far removed (From the strength of the material) must the use conditions be for many years of reliable operation depends on the degradation rate for the device/material. The stress and temperature dependence for this degradation rate is subject of reliability physics and is normally studied through the use of accelerated testing.

30.2 Accelerated Testing

By "accelerated testing," one means accelerating the normal TF process through the use of elevated stress and/or temperature. Accelerated testing is intended to shorten the normal TF process (which could take years), without changing the physics of failure. The objective of ULSI accelerated testing is to understand the stress and temperature dependence (kinetics) of failure mechanisms so that reliability improvements can be made. The reliability improvements are normally made through: defect elimination, lowering the intrinsic failure-rate (IFR), and making sure that wearout does not occur during the expected product lifetime. These three, distinctly different, reliability regions are generally displayed in a single reliability curve as shown in Figure 30.1. From its obvious shape, this reliability curve is commonly referred to as the "bathtub curve" for reliability.

The early failure-rate (EFR) portion of the reliability curve is dominated by defects (materials with extremely low breakdown strengths) and shows a sharp reduction in failure rate with time. Often the EFR will be inversely correlated with the yield (number of good die on the wafer compared with the total

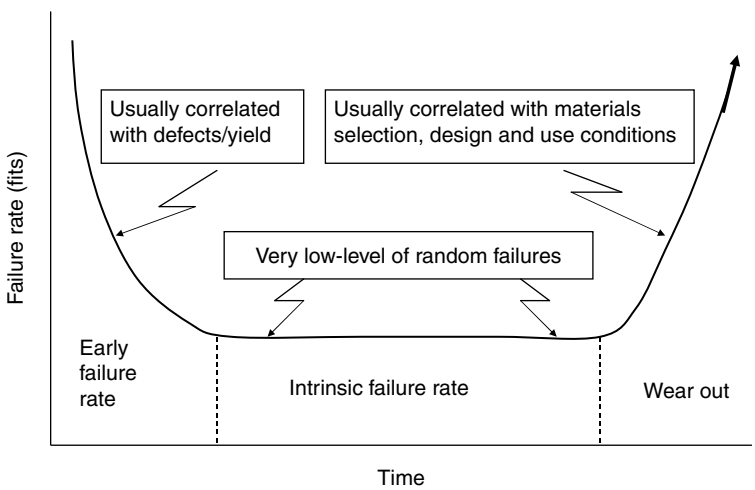


FIGURE 30.1 Bathtub curve can be used to describe the failure rate characteristics for ultra-large-scale-integrated-circuit devices.

number of die on the wafer). The failure rate that is observed at the bottom of the bathtub curve is nearly constant and is primarily due to those intrinsic weaknesses (intermediate breakdown strengths) found in a population of otherwise good devices; this portion of the curve we refer to as the IFR region. Finally, if the devices are operated long enough, they will eventually fail even though the material strength may be excellent; this region is referred to as the wearout region. The wearout region is correlated with the materials-type selection, design rules used, and the use conditions.

Accelerated testing is fundamental to ULSI reliability improvements because: (1) the defects (materials with low breakdown strengths) in a population of otherwise good devices can be eliminated with a short duration accelerated stress (burnin), (2) the intrinsic failures can be accelerated with stress/temperature (on a sampling basis) so that the IFR for a population of good devices can be determined, and (3) the time for wearout (TF for the main distribution of devices) can be projected from stress conditions to a specified set of operating conditions.

30.3 Time-to-Failure Modeling

All materials tend to degrade and will eventually wear out with time. For example, metals tend to creep and fatigue; dielectrics tend to lose their insulating properties and breakdown; paint tends to crack and peel; teeth tend to decay and fracture; etc. Ultra-large-scale-integrated-circuit devices also tend to degrade with time and eventually wear out. The rate of degradation and eventual TF will depend on the electrical, thermal, mechanical, and chemical environment to which the device is exposed.

In the case of metals, the extended nature of the valence-electron wave function for the metallic bond leads to bonding that is only weakly dependent on the exact location of individual metal-ions. [Note: this is the reason that metals can tend to show ductile and malleable properties.] For this reason, it is relatively easy for the metal ions to flow under the presence of an external force. While metal-ion movement is necessary for failure, it is not sufficient. For a material to degrade, and eventually fail, a flux divergence in the ion movement is required. By flux divergence, we mean that the flux of particles (number of particles per unit area per unit time) flowing into a region must be greater than or less than the flux of particles leaving the region.

A region of voiding or accumulation is depicted in Figure 30.2 as the result of a flux divergence in the particle transport process. This depiction could represent EM-induced voiding leading to an open circuit failure, a buildup of chlorine-ions on a bond pad leading to corrosion failure, or the trapping of electrons or holes in a dielectric leading to dielectric breakdown.

The flux divergence can be described by Fick's second law (which is a statement of the conservation of mass),

$$\vec{\nabla} \cdot \vec{J}(x,t) = -\frac{\partial \rho(x,t)}{\partial t} \tag{30.1}$$

where $J(x,t)$ represents the particle flux at the location x and time t, and $\rho(x,t)$ represents the density of such particles. Integrating both sides over the observation volume and using the divergence theorem, one

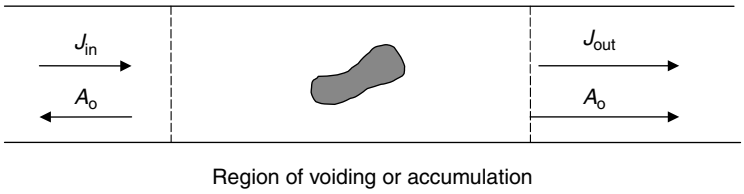


FIGURE 30.2 Material degradation (voiding and/or accumulation) occurs due to a flux divergence.

can express Equation 30.1 in integral form,

$$\int \vec{J} \cdot \mathrm{d}\vec{A} = -\frac{\mathrm{d}N(t)}{\mathrm{d}t} \tag{30.2}$$

where N represents the total number of particles contained in the volume of interest which is bounded by a closed surface of area A.

In analogy with reaction rate theory, it is convenient to think of the voiding (or accumulation) in terms of a reaction rate equation,

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = -k(t)N(t) \tag{30.3}$$

where $k(t)$ is the "reaction rate constant." Comparing Equation 30.2 and Equation 30.3, one obtains a relationship between the reaction rate constant and the flux divergence.

$$k(t) = \frac{\int \vec{J}(x,t) \cdot \mathrm{d}\vec{A}}{N(t)} \tag{30.4}$$

Failure is expected to occur when a given level of matter has depleted or accumulated in the failure volume. Using the general solution to Equation 30.3, one can define a TF as the time required for the number of particles to reach a critical depletion or accumulation level f_{crit},

$$\frac{N(t)}{N(0)} = \exp\left[-\int_0^t k(t')\mathrm{d}t'\right] \tag{30.5}$$

where

$$f_{\mathrm{crit}} = \frac{N(\mathrm{TF})}{N(0)} = \exp\left[-\frac{\int_0^{\mathrm{TF}} k(t')\mathrm{d}t'}{\int_0^{\mathrm{TF}} \mathrm{d}t'}\mathrm{TF}\right]. \tag{30.6}$$

Using Equation 30.4 and Equation 30.6, and solving for the TF, one obtains,

$$\mathrm{TF} = \frac{\ln(1/f_{\mathrm{crit}})}{\left\langle \frac{\int \vec{J}(x,t) \cdot \mathrm{d}\vec{A}}{N(t)} \right\rangle} \tag{30.7}$$

where the brackets $\langle\ \rangle$ refer to the time-averaged value of the quantities enclosed. The above equation shows explicitly that a flux divergence in the particle transport process is required to produce failure.

For many failure mechanisms, the transport of material can be described as Fickian-like,

$$J(x,t) = \mu\rho(x,t)F - D\frac{\partial\rho(x,t)}{\partial x} \tag{30.8}$$

where μ is the mobility and D is the diffusivity of the moving particles being acted upon by an average force F. The first term on the right-hand side of Equation 30.8 is referred to as the "drift component," while the second term is referred to as the "diffusion component." The relation between the mobility and

the diffusivity can be expressed by the Einstein relation,

$$\mu = \frac{D}{K_B T} = \frac{D_0 \exp\left(-\frac{Q}{K_B T}\right)}{K_B T} \tag{30.9}$$

where Q is the enthalpy of activation (usually referred to as simply activation energy), T is the temperature in degrees Kelvin, K_B is Boltzmann's constant (8.62×10^{-5} eV/K), and D_0 is the diffusion coefficient and is given by,

$$D_0 = \frac{\nu_0}{6} \tag{30.10}$$

where ν_0 is the vibration/interaction frequency $\sim 10^{13}$/s. Equation 30.7 through Equation 30.9 suggest that the TF should depend (exponentially) on temperature and on the driving force F.

It must be emphasized that even if we know the physics behind the driving force F and the activation energy for the diffusion process, which should permit accurate modeling of the flux given by Equation 30.8, seldom do we know the exact details of the flux divergence which is required in Equation 30.7 to model the TF precisely. For this reason, it is usually assumed that the flux divergence is related to the flux through either a power-law or exponential dependence. Thus, the TF Equation 30.7 is normally assumed to take one of the two forms:

$$\text{TF} = A_0(\xi)^{-n} \exp\left(\frac{Q}{K_B T}\right) \tag{30.11a}$$

or

$$\text{TF} = B_0 \exp(-\gamma \cdot \xi) \exp\left(\frac{Q}{K_B T}\right) \tag{30.11b}$$

In the above equations, ξ is the generalized stress (the agent which brings about the material degradation and eventual TF), n (or γ) is the stress dependent exponent, A_0 (or B_0) is a materials and process dependent coefficient, which is normally strongly microstructure dependent. The reliability physics parameters of primary interest are determined by accelerated testing and are obtained from Equation 30.11 by using:

$$n = -\left[\frac{\partial \ln \text{TF}}{\partial \ln \xi}\right]_T \tag{30.12a}$$

or

$$\gamma = -\left[\frac{\partial \ln \text{TF}}{\partial \xi}\right]_T \tag{30.12b}$$

and

$$Q = K_B \left[\frac{\partial \ln \text{TF}}{\partial (1/T)}\right]_\xi \tag{30.13}$$

In Figure 30.3 and Figure 30.4 we illustrate how the above kinetics can be obtained from observed TF data. For this illustrative example, the stress dependence for the TF was described by a power-law with exponent $n = 2$, and the temperature dependence was described as being Arrhenius-like with an activation energy of 0.5 eV.

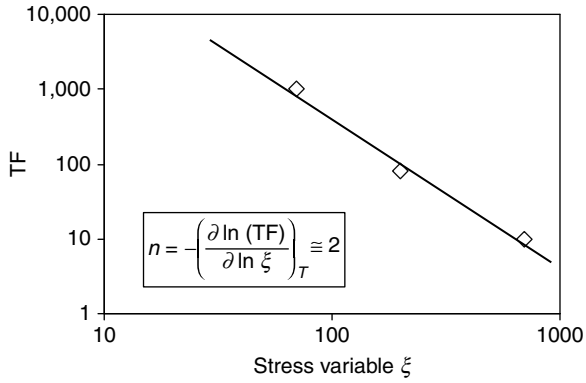


FIGURE 30.3 Determination of the stress dependence for time-to-failure.

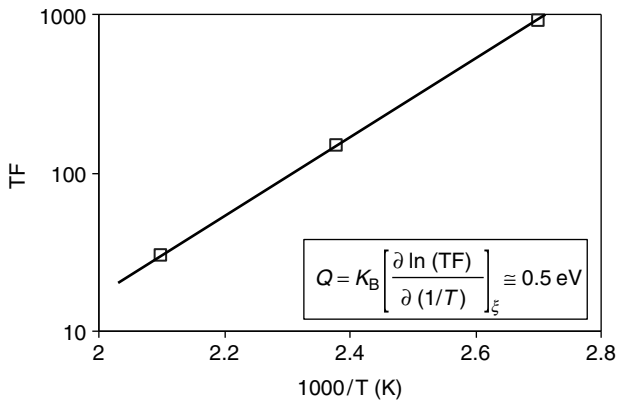


FIGURE 30.4 Determination of the activation energy for time-to-failure.

It must be emphasized that when determining the stress dependence n (or γ), using Equation 30.12, the temperature must be held constant (either physically or mathematically) while the stress level is being changed. [Note: this can be a very important issue if the stress also tends to serve as a significant additional source of heating, e.g., joule heating can raise the temperature of the sample when the current density stress is increased in a metal stripe during EM testing.]

Therefore, one may wish to determine the activation energy first, using Equation 30.13, with the stress level held fixed, and then use the activation energy to derate to a fixed temperature as the current density is varied for n determination according to Equation 30.12a.

30.4 Time-to-Failure Statistics

When identically processed materials are placed under the same level of stress, they will not fail exactly at the same time. (An explanation for this occurrence is that a slight difference exists in breakdown strengths even for identically processed material; i.e., the process dependent coefficients A_0 and B_0 in Equation 30.11 can vary from sample to sample.) This means that not only are we interested in TF, but more precisely, we are interested in the distribution of times-to-failure. Once the distribution of times-to-failure is established, then one can construct a probability density function $f(t)$ which will permit one

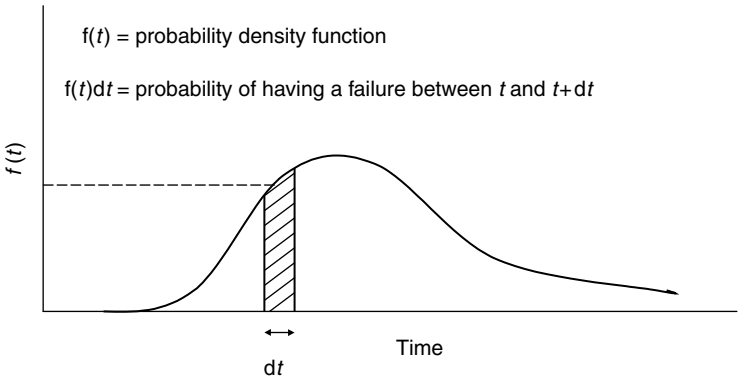


FIGURE 30.5 Probability density function $f(t)$ for failures.

to estimate the probability of observing a failure in any arbitrary time interval between t and $t+\mathrm{d}t$, as is illustrated in Figure 30.5.

Historically, three probability density functions have been widely used to describe very large scale integration (VLSI) failures: exponential, Weibull, and lognormal. Since the exponential is a special case of the Weibull, only the latter two distributions will be discussed here in some detail.

30.4.1 Lognormal Distribution

The lognormal distribution is based on the normal distribution except that failures are assumed to be logarithmically distributed in time rather than linearly distributed. The lognormal probability density function is defined by,

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} \exp\left\{-\left[\frac{\ln(t) - \ln(t_{50})}{\sigma\sqrt{2}}\right]^2\right\} \tag{30.14}$$

where t_{50} is the median TF and σ is the logarithmic standard deviation (SD). σ is usually approximated by $\sigma = \ln(t_{50}/t_{16})$ where t_{16} represents the TF for 16% of the units.

Both t_{50} and t_{16} can be read directly from a cumulative failure probability $F(t)$ plot with lognormal scaling, where the cumulative failure probability $F(t)$ is defined as:

$$F(t) = \int_0^t f(t')\mathrm{d}t' \tag{30.15}$$

To better illustrate this, the cumulative failure data found in the Table 30.1 has been plotted with lognormal scaling and is shown in Figure 30.6.

The use of the lognormal distribution has been very popular for ULSI failure mechanisms such as EM failures (which will be discussed in more detail later in this chapter) [4,5]. A lognormal plot can be an effective way to see if a single or multiple populations are present. Multiple populations are indicated if: (1) there is a substantial curvature to the plotted data (poor correlation coefficient) or (2) there is a substantial change in slope in the data. If multiple populations are present, one must separate the data such that each distribution contains a single "pure" mechanism. This will be indicated by an improvement in correlation coefficient and no abrupt change in slope of the data.

TABLE 30.1 Cumulative Time-to-Failure Data

Read Points (h)	Cumulative % Failures
500	2
1000	22
1500	50
2000	72

30.4.2 Weibull Distribution

The Weibull distribution is a "weakest link" type distribution. By using the term weakest link one means that the failure of the whole (chain) is dominated by the degradation rate for the weakest element (link). The Weibull probability density function is defined by

$$f(t) = \left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \tag{30.16}$$

where α is the characteristic TF and β is referred to as the shape (or dispersion) parameter. The values for α and β can be determined from a cumulative failure probability plot on Weibull probability paper. The following equations can sometimes be useful when working with Weibull plots:

$$\beta = \frac{1.38}{\ln(t_{50}/t_{16})} \tag{30.17}$$

and

$$\alpha = \frac{t_{50}}{[\ln(2)]^{1/\beta}} \tag{30.18}$$

Unlike the lognormal distribution (where the cumulative failure probability $F(t)$ must be obtained by numerical methods), an analytical expression can be found for the cumulative Weibull failure probability

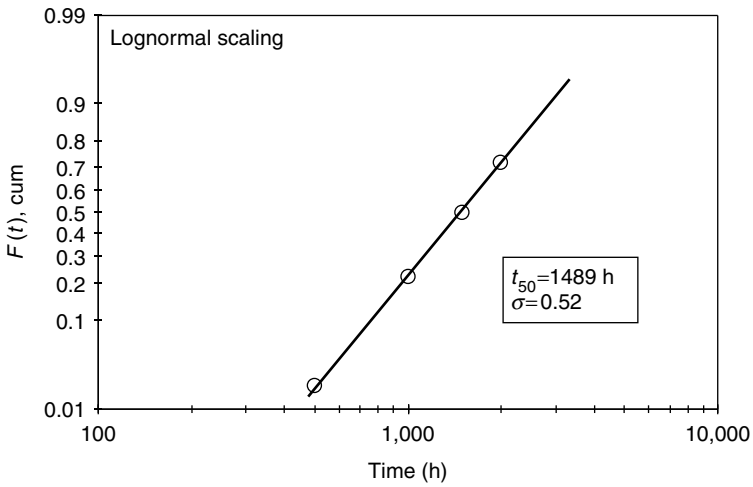


FIGURE 30.6 Lognormal plotting of cumulative data from Table 30.1.

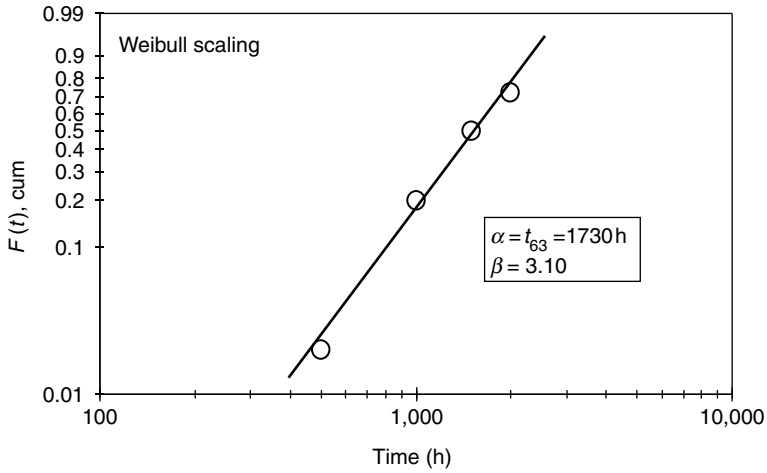


FIGURE 30.7 Weibull plotting of cumulative data from Table 30.1.

function,

$$F(t) = \int_0^t f(t')\mathrm{d}t' = 1 - \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \tag{30.19}$$

A Weibull plot of the cumulative failure data found in Table 30.1 is shown in Figure 30.7.

An alternative, and currently popular, way of displaying Weibull cumulative failure data is realized by rewriting Equation 30.19 in the form:

$$\ln[-\ln(1-F)] = \beta \ln(t) - \beta \ln(\alpha) \tag{30.19a}$$

This alternative way of performing the Weibull plotting is shown in Figure 30.8 for the cumulative failure data found in Table 30.1 [Note that the left hand side of Equation 30.19a goes to zero when F is approximately 0.63 showing that the characteristic time α is approximately equal to t_{63}. Also note that, according to Equation 30.19a, β is simply the slope of the line in Figure 30.8.].

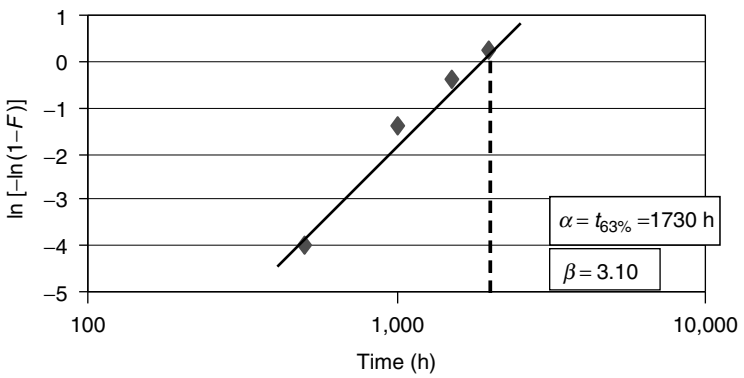


FIGURE 30.8 Alternative method for Weibull plotting of data from Table 30.1.

The Weibull distribution seems to be the preferred distribution for plotting ULSI failure mechanisms such as time-dependent dielectric breakdown (TDDB) data [6], an important ULSI failure mechanism which will be discussed in more detail shortly. The Weibull distribution tends to fit TDDB data extremely well because one small localized region of the dielectric will tend to degrade much more rapidly than the other regions of the dielectric. Thus, the failure of the whole (capacitor) tends to be dominated by the degradation of this weakest link (localized region in the dielectric).

30.5 Failure Rate

The failure rate determination for a collection of ULSI chips is of primary reliability importance. The rate equation, by which the devices are expected to fail, is given by

$$\frac{dM}{dt} = -\lambda(t)M(t) \tag{30.20}$$

where $M(t)$ represents the number of survivors at any time t and $\lambda(t)$ is referred to as the survivor failure rate. Since the number of surviving units at any time can be expressed by

$$M(t) = M(0)[1 - F(t)] \tag{30.21}$$

then the failure rate λ is given by:

$$\lambda(t) = -\frac{1}{M(t)}\frac{dM}{dt} = \frac{f(t)}{1 - F(t)} \tag{30.22}$$

The accepted unit for failure rate is the fit where 1-fit represents 1-failure per billion device·h. For example, if a system is made up of 1000 chips, with each chip having an IFR of 100 fits, then one would expect a system failure every 1.14 year, e.g.,

$$\text{System failure rate} = 1000 \text{ devices} \times \left(\frac{100 \text{ failures}}{10^9 \text{ device·h}}\right) = \frac{1 \text{ failure}}{10^4 \text{ h}} \tag{30.23}$$

By using either the lognormal probability density function (Equation 30.14) or the Weibull probability density function (Equation 30.16), in conjunction with Equation 30.22, the failure rate can easily be determined under accelerated test conditions. To go from the empirically determined failure rate under accelerated conditions to a projected failure rate under normal operating conditions, the concept of an acceleration factor must be introduced.

30.6 Acceleration Factor

The concept and use of an acceleration factor is of great and fundamental importance to the theory of accelerated testing. The acceleration factor AF is defined as the ratio of the expected TF under normal operating conditions to the TF under accelerated stress conditions,

$$\text{AF} = \frac{(\text{TF})_{\text{operation}}}{(\text{TF})_{\text{stress}}} \tag{30.24}$$

Using the TF model, Equation 30.11, one obtains,

$$\text{AF} = \left(\frac{\xi_{\text{stress}}}{\xi_{\text{op}}}\right)^n \times \exp\left[\frac{Q}{K_B}\left(\frac{1}{T_{\text{op}}} - \frac{1}{T_{\text{stress}}}\right)\right] \tag{30.25}$$

In the last equation, if the values of n and Q are determined from accelerated testing conditions using Equation 30.12 and Equation 30.13, one can then easily model the acceleration factor. Note that the process/materials-dependent prefactor A_0, which appears in and greatly impacts the TF in Equation 30.11, does not appear in the expression for the acceleration factor. This means that the acceleration factor depends only on the physics of failure (kinetics) and not on how good or how bad the material is relative to its time to failure.

The modeled acceleration factor, Equation 30.25, permits one to go from the failure rate data taken under accelerated test conditions to the projected failure rate under normal operating conditions. The transformations from stress conditions to operating conditions for the lognormal and Weibull distributions are given by:

lognormal

$$\left(t_{50}\right)_{\text{op}} = \text{AF} \cdot \left(t_{50}\right)_{\text{stress}} \tag{30.26a}$$

and

$$(\sigma)_{\text{op}} = (\sigma)_{\text{stress}} \tag{30.26b}$$

Weibull

$$(\alpha)_{\text{op}} = \text{AF} \cdot \alpha_{\text{stress}} \tag{30.27a}$$

and

$$\beta_{\text{op}} = \beta_{\text{stress}} \tag{30.27b}$$

Please note that while the characteristic TF for each distribution has been transformed using the acceleration factor, it has been assumed that the dispersion parameter (α,β) for each distribution does not change with stress. This will serve as the definition for uniform acceleration, i.e., uniform acceleration tends to accelerate the entire TF distribution uniformly such that the dispersion/slope of the distribution does not change with the level of stress. One should always take enough accelerated data to establish under what set of stress conditions is the acceleration uniform such that the constant dispersion parameter assumption is valid. A change in slope of the TF distribution with stress level may indicate a change in physics could be occurring; i.e., one is not simply accelerating the physics of failure but one is changing the physics of failure. The goal of accelerating testing is to accelerate the physics without changing the physics.

30.7 Time-to-Failure Models for Selected ULSI Failure Mechanisms

30.7.1 Electromigration

Electromigration has historically been a significant reliability concern for Al-based metallizations [7–21]. Due to the momenta exchange between the current carrying electrons and the host metal lattice, aluminum-ions will drift under the influence of the electron wind. Equation 30.8 can be used to describe the flux J of metal ions where the force F on the metal ions due to the electron wind is given by,

$$F = R_0 z^* e J^{(\text{e})} \tag{30.28}$$

where R_0 is the metal resistivity, z^*e is the effective ion charge, and $J^{(\text{e})}$ is the electron current density. Eventually, due to a flux divergence (caused by gradients in microstructure, temperature, stress,

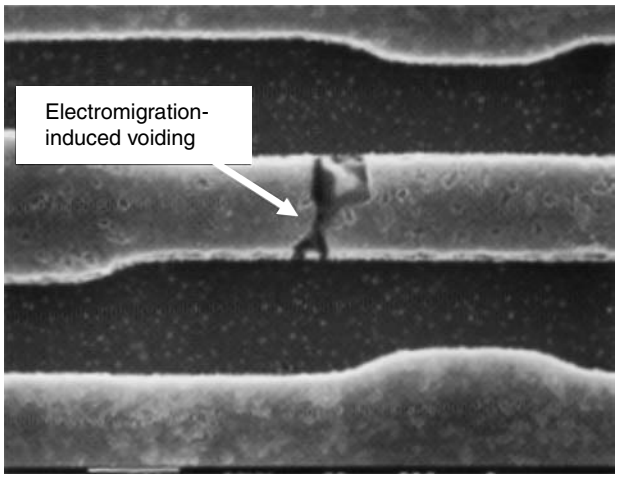


FIGURE 30.9 Electromigration (EM)-induced transport (and eventual flux divergence) has produced the severe voiding in the Al metal lead shown.

impurities, etc.) vacancies will start to cluster, the cluster will grow into a void, and finally the void growth will continue until the conductor reaches a resistive or open circuit condition (see Figure 30.9).

For Al-alloys, the metal-ion transport is primarily along grain boundaries, as illustrated in Figure 30.10. The void nucleation phase generally has little/no impact on the electrical resistance rise of the metal stripe. The void growth phase, however, can cause local current crowding and a rise in resistance. If the metallization is actually an Al-alloy/barrier-metal laminate, then the resistance rise may be gradual as illustrated in Figure 30.11. This gradual rise in resistance, of course, assumes that the barrier metal is EM resistant. Some commonly used EM resistant barriers in ULSI applications include: TiW, TiN, and TaN. Without a barrier layer present to participate in carrying the current, the rise in resistance for the Al-alloy can be very abrupt for EM- induced damage.

For pure copper metallization, the dominant diffusion path during EM testing is reported to be along interfaces, not along grain boundaries [22]. In order to prevent the Cu-drift, the Cu lead must be bounded fully by barrier layers. Normally, the bottom and sidewalls of the Cu lead are protected by a TiN

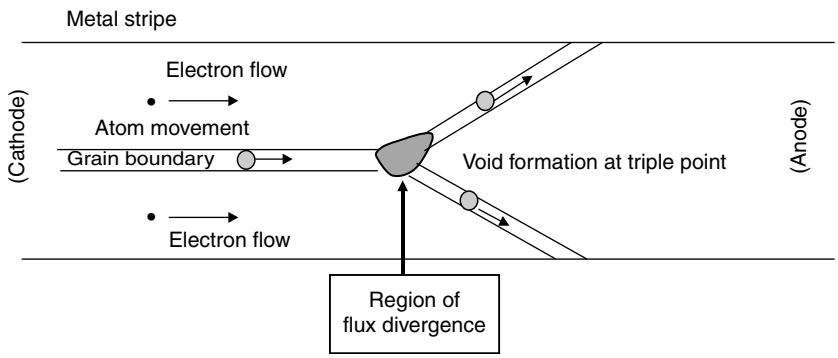


FIGURE 30.10 EM transport is primarily along grain boundaries in polycrystalline Al conductors. Flux divergence at a grain-boundary triple point can produce voiding or accumulation. The dominant EM transport mechanism for pure Cu may be along the Cu/barrier interfaces.

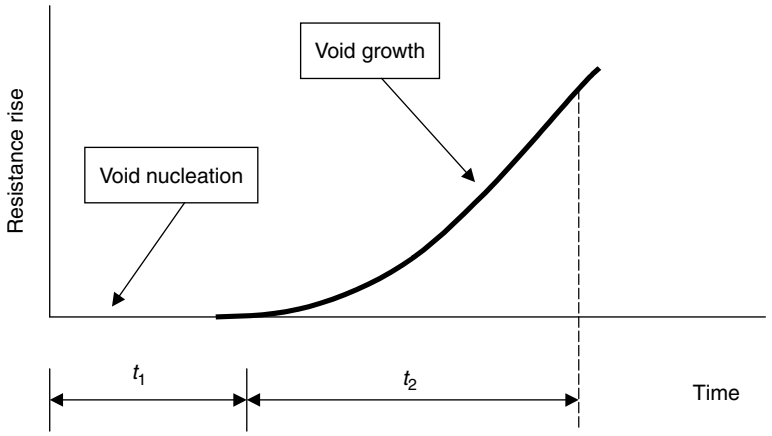


FIGURE 30.11 EM-induced resistance rise in layered metal stripes (e.g., Al–Cu/TiN) shows little/no resistance rise initially and then a gradual resistance rise. The TiN layer serves as an EM-resistant shunting layer to prevent catastrophic opens.

or TaN barrier while the top of the Cu lead has a dielectric barrier. During EM transport, the Cu will select one or more of the interfaces along which it can diffuse with greater mobility.

While differences in the materials properties between Cu and Al lead to a different dominant mass transport mechanism, Cu metallization is also distinguished from Al because it is fabricated differently using the so-called "damascene" or "dual-damascene" process flow (See Figure 30.12), rather than the deposition and subtractive etch process used to make earlier technology Al-based interconnects [23,24].

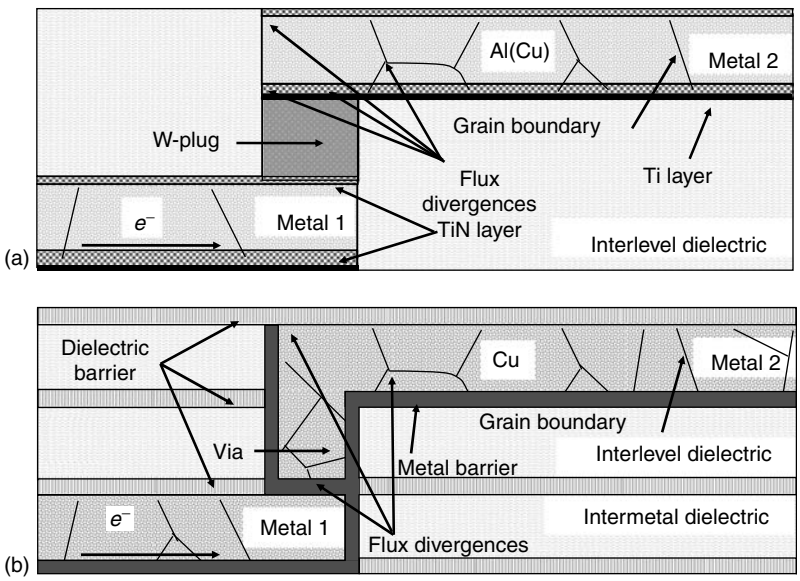


FIGURE 30.12 Flux-divergence locations for (a) Al-based metallizations and (b) Cu-based metallizations. Thicknesses and aspect ratios are not to scale.

The change in process methodology is primarily dictated by the lack of effective Cu etching processes. In the damascene flow, trenches are first fabricated into a dielectric layer (where the metallization will eventually go) and then conformally lined with a metal barrier material such as Ta-based metallization and a thin physically vapor-deposited Cu seed. This trench feature is then filled with Cu metallization using an electroplating process (EP). This is followed by chemical mechanical polishing (CMP), and subsequently cleans to define the interconnect geometry. Then, the layer is capped by a sealing layer, usually a dielectric barrier material such as SiN_x or SiC_xN_y.

In the dual-damascene process, the via openings are also formed in addition to the trench such that a via and trench are not separated by a metal barrier as would be the case for single damascene interconnects. In dual-damascene Cu, a flux barrier (due to the metal barrier) is present at the via bottom. This somewhat complicated interconnect architecture, utilizing dielectric and metal barriers with different interface properties, exhibits a number of flux divergence locations not seen in Al-metallization [25]. For electron flow up into a via (up-direction EM), a flux divergence is located at the top corner of the trench, but for down-direction EM, the flux divergence location is located along the top surface of the lower metal trench where the metal barrier of the via and the dielectric cap on the lower metal trench meet. The voiding volumes necessary to cause circuit open are also somewhat different for the two cases—leading to the general observation that down-direction EM failures occur a bit faster than up-direction EM. Additionally, defects present within a via may lead to premature EM failure (early or weak-mode failure) of an up-direction interconnect [23].

The presence of weak interfaces in Cu metallizations, due to the fact that Cu does not form a stable native oxide, means that optimization of interfacial adhesion strengths between Cu and the capping layer are critical [26]. Studies have shown that improvements in interfacial adhesion scale well with improvements in EM performance [27]. Also, where the interface quality is rendered exceedingly high as in the case of a Co-based metal cap on the Cu trench, the EM performance becomes limited by the bulk properties of the Cu metallization [28].

Since EM transport is a mass conserving process then, in addition to the voiding problems, accumulations of the transported metal ions will also occur and may severely increase the mechanical stress in surrounding dielectrics. This buildup of mechanical stress can serve to locally increase the mechanical stress and serve to generate a "back-flow" of metal ions (the Blech effect, Ref. 10). For short leads (generally a few microns), the Blech effect can be so strong that the back flow will cancel the drift component in Equation 30.8 and EM failure can be retarded. However, the buildup of mechanical stress in the metal lead is also accompanied by a buildup of mechanical stress in the surrounding dielectrics causing fracture. Fracture of the surrounding dielectrics can facilitate the shorting of the test lead to the adjacent metal leads. For advanced Cu metallizations that require low-k dielectrics, that are generally weak mechanically, this sort of failure mechanism has become an increasing concern [29–31].

The model generally accepted to describe EM TF takes the form [7–21]:

$$\mathrm{TF} = A_0 (J^{(e)} - J^{(e)}_{crit})^{-n} \exp\left(\frac{Q}{K_B T}\right) \tag{30.29}$$

where

A_0, process/material-dependent constant;
$J^{(e)}$, electron current density [Note: $J^{(e)}$ must be greater than $J^{(e)}_{crit}$ to produce failure];
$J^{(e)}_{crit}$, critical (threshold) current density which must be exceeded before failure will occur. $J^{(e)}_{crit}$ is related to the Blech length for the lead being evaluated;
$J^{(e)}_{crit} \cdot L_{crit} \sim 6000$ A/cm, for Al-alloys and ranges from 4000 (oxide dielectric) down to nearly 1000 A/cm (porous low-k), depending on the mechanical strength of the surrounding dielectric and metal barrier materials [29–33]. If the test stripe length is >250 μm, then J_{crit} is small compared to the normal EM stressing current density of >1 MA/cm^2;

n, current density exponent. [Note: $n=2$ for incubation period or $n=1$ for resistance-rise period in layered metal systems]; and
Q, activation energy ($Q=0.5$–0.6 eV for Al and Al–Si, $Q=0.7$–0.9 eV for Al–Cu alloys, and $Q=0.9$ eV or greater for pure Cu).

Electromigration data is normally collected under dc conditions whereas the circuit operation is ac. This means that a method is needed to transform ac current densities into dc EM equivalents for design rule checking. For unipolar current waveforms, $J^{(e)}$ can be taken as the average current density $\langle J^{(e)}\rangle$ [18,19]. For bipolar current waveforms, a "sweep-back" recovery action can take place and the effective current density $J^{(e)}$ can be described by $J^{(e)} = \langle J_+^{(e)}\rangle - r\langle J_-^{(e)}\rangle$, where $\langle J_+^{(e)}\rangle$ is the average of the positive polarity pulses and $\langle J_-^{(e)}\rangle$ is the average of the negative polarity pulses. "r" is the recovery coefficient and has a value of at least 0.7.

Electromigration associated with vias must be investigated separately because they show characteristics which are different from single leads fed by bonding pads. For example, vias can show different degradation rates depending on electron current flow direction (upper-level of metal M2 to lower-level of metal M1 may be quite different from M1 to M2). Also, the degradation rate is strongly dependent on via structure (barrier layer, capping layer, and via-etching), via number and layout, and a reservoir effect can be present [20,21].

For Al-alloy stripes, terminated by bonding pads and having no barrier metallization, the total TF is dominated by nucleation, and n is observed to be equal to 2 (which is commonly referred to the Black equation [8]). However, for Al-alloy stripes with barrier metal terminated by tungsten plugs, one may see both an incubation (nucleation) period, dominated by $n=2$ but and a resistance rise (drift period) dominated by $n=1$ (as is illustrated in Figure 30.11). Also, under high current density test conditions, unaccounted for Joule-heating can produce apparent current density exponents much greater than 2. Similar observations hold for Cu metallization, where a mixture of nucleation and growth contributions is often simultaneously present; however, the trend appears to be weighted more towards growth-controlled EM [34]. Thus, caution must be exercised when extrapolating TF data from high to low current densities.

30.7.2 Contact Electromigration

Ultra-large-scale-integrated-circuit metallization must be used to make contact to shallow (<0.25 um) N^+ and P^+ junctions in Complimentary metal oxide semiconductor (CMOS) technologies. Being able to build a stable/reliable contacts necessitates that a barrier metal be used between the interconnect metal (either Aluminum-alloy, W-plug or Copper) and the shallow junction. Some common barrier metals often used are TiW, TiN, and TaN. During contact EM transport, the dominant diffusing species which causes contact failure is reported to be silicon from the contact region [14–17]. In addition to the barrier type being important, silicided junctions can also be important relative to the transport process.

Equation 30.29 can also be used to describe ULSI contact (metal to silicon or silicide) failure due to EM [17]. Here, however, the diffusing species which leads to failure is generally the silicon. Contact EM failure occurs when a buildup of silicon occurs in the contact window leading to resistive contact formation; or, an erosion of silicon from the contact window can lead junction leakage and failure. Since the current crowding can be severe in a shallow contact, the actual current density is non-uniform over the contact window and may be very difficult to specify. For this reason, normally the contact area is incorporated into the process dependent prefactor A_0 and the TF equation becomes:

$$\mathrm{TF} = A_0 I^{-n} \exp\left(\frac{Q}{K_B T}\right) \tag{30.30}$$

where I is the current flowing into or out of the contact window. For aluminum-alloy to silicon contacts, the reported values of activation energy are generally in the range 0.8–0.9 eV while for silicided ($TiSi_2$,

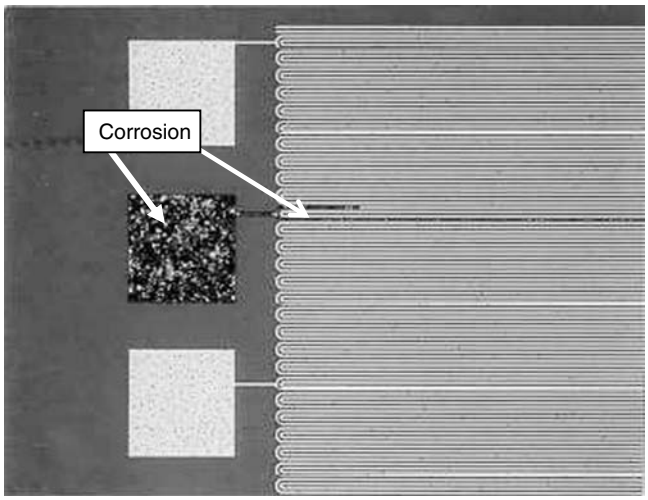


FIGURE 30.13 Corrosion of aluminum bonding pads can occur if chlorides and moisture are present.

$TaSi_2$) contacts the values are higher 1.1–1.5 eV [14–17]. Due to the extreme localized nature of the self-heating during contact stressing, the reported values for n can vary greatly 1–11 [17].

30.7.3 Corrosion

Corrosion failures can occur when ULSI devices are in the presence of moisture and contaminants [35–44]. Corrosion failures are usually classified as one of two broad groups: bonding-pad corrosion or internal-chip corrosion. Bonding-pad corrosion (see Figure 30.13) is usually more common simply because the die passivation does not cover the metallization in the bonding pad locations. Internal corrosion (internal to the chip, away from the bonding pads) can also occur if some weakness or damage exists in the die passivation which would permit the moisture and contaminants (e.g., chlorine) to reach the metallization.

Corrosion can generally be described in terms of a "corrosion cell" where there must exist four key components in order for corrosion to occur: an anode, a cathode, an electrolyte, and an conductor to provide a path for the electron-flow needed for the oxidation/reduction processes. An example of "wet" corrosion is shown in Figure 30.14. Metal corrosion (oxidation) can occur if there is an imperfection in

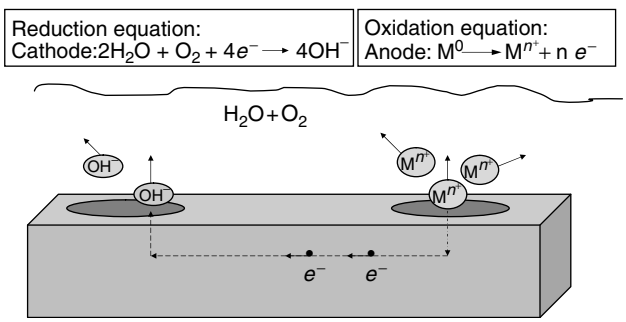


FIGURE 30.14 "Wet" corrosion generally occurs with low activation energy because of the very high mobility of the diffusing species in water.

the native oxide covering the metallization. Generally, Al forms a good self-passivating oxide and is much less corrosive than Cu, even though the Galvanic series would suggest just the opposite. However, if chlorine is added to the water, then the Al_20_3 native oxide protecting the Al will quickly be reduced, exposing a highly reactive virgin Al surface which will then rapidly corrode [40].

In order for the corrosion to continue at a rapid rate, the contaminants and metal ions must be able to diffuse rapidly to and from the corroded region, respectively. This can occur easily in liquids and the activation energy for liquid/wet corrosion is generally very low ~0.3 eV. However, for "dry" or "ambient" corrosion (see Figure 30.15) the activation energy for diffusion is generally higher and the corrosion rate is very dependent on the percent relative humidity (%RH). In fact, the surface mobility has been shown to be exponentially dependent on %RH over a rather wide range of %RH [37].

Three industry-standard tests have been widely used to accelerate potential ULSI corrosion failure mechanisms: biased 85°C and 85%RH, autoclave (121°C and 100%RH), and highly accelerated stress test (HAST) conditions (typically: biased, 121°C, 85%RH). To extrapolate accelerated corrosion results to field use-conditions [from relative humidity (RH) and absolute temperature], at least three models have been reported and used:

Reciprocal Exponential Humidity Model [41]:

$$\mathrm{TF} = A_0 \exp\left(\frac{b}{\mathrm{RH}}\right) \exp\left(\frac{Q}{K_\mathrm{B} T}\right) \tag{30.31a}$$

where A_0, Process/materials dependent parameter; b, Reciprocal humidity dependence parameter (~300%); RH, Relative Humidity expressed as % (Note: 100% = saturated); Q, 0.3 eV (Note: typical for aluminum corrosion with phosphoric acid present).

Power-law humidity model [42]:

$$\mathrm{TF} = A_0 (\mathrm{RH})^{-n} \exp\left(\frac{Q}{K_\mathrm{B} T}\right) \tag{30.31b}$$

where n, ~2.7; RH, Relative Humidity expressed as % (Note: 100% = saturated); Q, 0.7–0.8 eV (Note: typical for aluminum corrosion when chlorides are present).

Exponential humidity model [37–44]:

$$\mathrm{TF} = A_0 \exp(-a \cdot \mathrm{RH}) \exp\left(\frac{Q}{K_\mathrm{B} T}\right) \tag{30.31c}$$

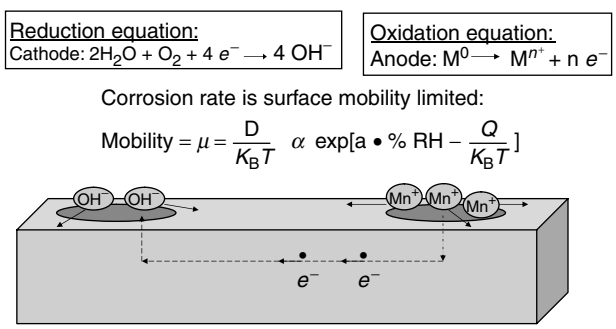


FIGURE 30.15 "Dry" or ambient corrosion is strongly humidity-dependent because the percent relative humidity (%RH) greatly impacts surface/interface mobility of the diffusing species.

where a, [0.10–0.15 $(\%)^{-1}$]; RH, Relative humidity expressed as % (100% = saturated); Q, 0.7–0.8 eV (Note: typical for aluminum corrosion when chlorides are present).

There seems to be reasonably good consensus that the proper activation energy for chlorine-induced aluminum corrosion is in the 0.7–0.8 eV range. There is not a consensus for the humidity dependence. A fairly recent comparison of the three models [44] tended to favor the exponential model with a $\sim$[0.12–0.15 $(\%)^{-1}$]. However, currently, the power-law model is probably the most widely used corrosion model in the industry.

30.7.4 Stress Migration

Mechanical stress related failures are very important for ULSI devices [45–50]. When a metal is placed under a mechanical stress which exceeds its yield point, the metal will undergo plastic deformation with time. This time-dependent phenomenon is described by metallurgists as creep. The creep will continue until the stress level is brought below the yield point or until the metal fails. This metal failure mechanism is especially important for ULSI where one is confronted with: on-chip aluminum-alloy or copper metallization, gold ball-bonds and wires, iron-alloy or copper lead frames, solder joints, etc.

Stress migration (SM) is the term used to describe the flow of metal atoms under the influence of mechanical-stress gradients. Generally, stress gradients can be assumed to be proportional to the applied mechanical stress σ (residual thin film or thermomechanical). Relatively little atom movement (plastic flow if "cold" or creep if "hot," i.e., hot means $T \sim T_{\text{melting}}/2$) occurs until the stress σ exceeds the yield-point of the metallization. The flux of the moving metal atoms is primarily along grain boundaries as is illustrated in Figure 30.16, but may also occur within a grain if the metal lead is very narrow and the grain structure can be considered "bamboo."

The inevitable flux divergence associated with the metal movement will cause notching and voiding in the ULSI metal leads/stripes to occur (see Figure 30.17). The resistance rise associated with the void formation can cause electrical failures [45–50]. The TF has been described by:

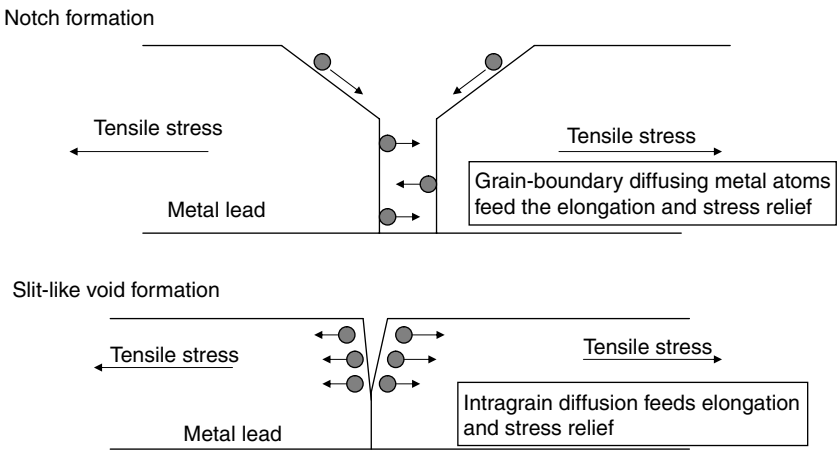


FIGURE 30.16 Mechanical-stress gradients can cause metal atoms to flow in an effort to relieve the stress. For fine-grain Al (mean grain size < metal strip width), diffusion is generally along grain boundaries. For coarse-grain Al (mean grain size > metal stripe width which produces a "bamboo- like" grain structure), the dominant diffusion path may occur within a grain. As for Cu, the dominant diffusion path may be along Cu/barrier interfaces.

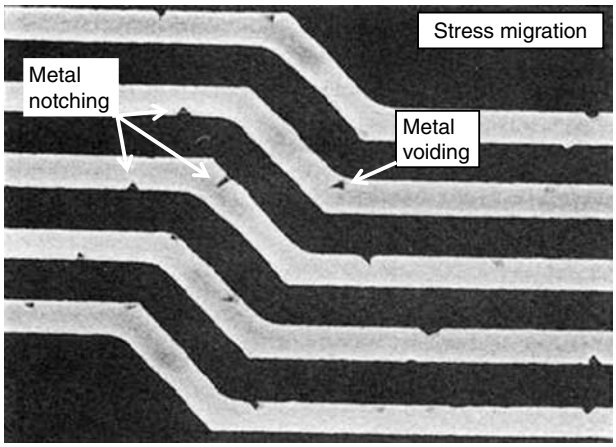


FIGURE 30.17 Stress migration has served to produce notching and voiding in the Al-alloy metal leads shown.

$$\mathrm{TF} = A_0 \sigma^{-n} \exp\left(\frac{Q}{K_B T}\right) \tag{30.32}$$

where σ, constant stress load; n, 2–3 for soft metals; Q, activation energy $\cong$0.5–0.6 eV for grain boundary diffusion, $\sim$1 eV for single grain (bamboo-like) diffusion.

If the mechanical stress is generated by thermal expansion mismatch of a "stack" of thin film materials, then the stress is referred to as "thermomechanical stress" and σ is proportional to the change in temperature, i.e.,

$$\sigma \propto \Delta T. \tag{30.33}$$

Therefore, if the metal creep is caused by thermomechanical stress, then the TF can be expressed by the McPherson and Dunn model for SM [47],

$$\mathrm{TF} = A_0 (T_0 - T)^{-n} \exp\left(\frac{Q}{K_B T}\right) \tag{30.34}$$

where T_0, stress-free temperature for metal.

The role of stress and stress relaxation is very important in the nucleation and growth of voids in aluminum-alloy interconnects. Cu doping in the aluminum is somewhat effective in suppressing grain-boundary diffusion, but is much less effective if the grain size is large compared to linewidth. In these "bamboo-like" leads, one observes slit-like void formation due to intra-grain diffusion.

Currently, there is no industry accepted standard test for stress migration (SM). Typically, long (>1000 μm) and narrow (<2 μm width) stripes are stored at temperatures in the range 150°C–250°C for 1–2 Kh and then electrically tested for resistance increases or reduction in breakdown currents. The SM baking temperature should be carefully selected because, as predicted from Equation 30.34, there is a maximum in the creep rate (as illustrated in Figure 30.18). This maximum generally occurs in the 150°C–250°C range and drives a minimum in the TF (Equation 30.34). This maximum in the creep rate occurs due to the high-stress but low-mobility at lower temperatures, and low-stress but high-mobility at high temperatures [47].

Since the mechanical stress is temperature dependent, a straightforward determination of the diffusion activation energy is somewhat difficult to obtain. Generally, however, $Q \sim$0.5–0.6 eV is used for grain-boundary diffusion and $\sim$1 eV for single-grain (bamboo-like) diffusion.

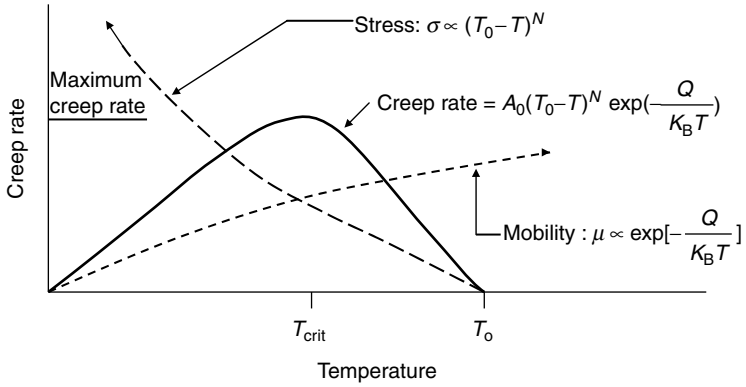


FIGURE 30.18 Stress-migration induced degradation has a maximum at a critical temperature (which is generally in the 150°C–250°C range for Al-alloys). This maximum in the damage (creep) rate occurs because of the low mobility (but high stress) at lower temperatures and low stress (but high mobility) at elevated temperatures.

The use of refractory metal barriers or layered metallization has tended to greatly reduce the impact of the damage caused by slit-like void formation in bamboo leads because the refractory metal layer tends to serve as a redundant conductor, shunting the current and reducing the electrical resistance rise when a SM-induced void forms.

Stress migration in Cu metallization is also a concern [51–56] despite an expectation that Cu's generally superior EM capabilities would translate to significantly improved SM performance (see Figure 30.19). Similar to a comparison between Al and Cu EM, as it pertains to its use in advanced integrated circuit technology, the contrasting fabrication methods for Cu versus Al lead to pronounced differences in the type of SM issues found in the different metallizations. A basic difference between Cu and Al lies with their different melting points: 1083 vs. 660°C. Normal processing temperatures during

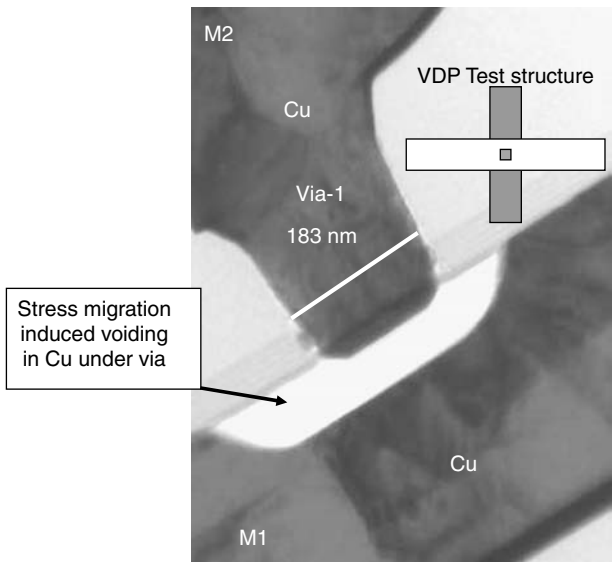


FIGURE 30.19 Stress-induced voiding under a Van der Pauw (VDP) via in a copper interconnect system.

integrated circuit fabrication can be as high as 400°C, which is a substantial fraction of the melting temperature of Al but not Cu. Hence, processing of Al metallization can lead to grains that are large and well-formed within interconnect wiring (so-called bamboo structure), but similar processing temperatures do not greatly alter the microstructure of Cu that has reached a certain level of stability. So, the grain structure within Cu interconnect wiring is much more varied, both grain size-wise and texture-wise [51]. Electroplated Cu also greatly impacts the evolved microstructure such that narrow lines remain small grained whereas wider lines develop larger grains from an initially small average size but only after sufficient heat treatment or self-annealing [52].

Like Al lines, Cu lines can show evidence of voiding; however, because of the presence of a somewhat redundant metal barrier and the lack of sufficient bamboo character to the Cu grains, its impact on reliability is not so strong [53]. When a void is found under or within vias as shown in Figure 30.19, the reliability impact can be substantial, especially because a via is electrically a weak-link along the interconnect path. This impact is felt most severely when wide leads are placed over or under single vias [54,55]. Once a void is nucleated, an ample supply of vacancies can be provided within wide leads to enlarge the void within or under a via and enable open circuits to form [54,56].

30.7.5 Cyclic Fatigue

Fatigue failures can occur in ULSI devices due to thermal cycling [57–62]. For example, the lifted bonding ball shown in Figure 30.20 resulted from the thermomechanical stress during temperature cycling. The thermomechanical stress (generated by the thermal expansion coefficient mismatch of the: plastic molding compound, gold bonding ball, Au–Al intermetallics, Al pad, and silicon substrate) served to weaken the bond during each thermal cycle and eventually led to failure of the bond.

Thermal cycling of a device will naturally occur each time the assembled chip undergoes a normal power-up and power-down cycle. Such thermal cycles can induce a cyclical thermomechanical stress that tends to weaken the materials [57–62], and may cause: dielectric/thin-film cracking, lifted bonds, fractured/broken bond wires, solder fatigue, cracked die, etc.

The thermomechanical stresses during thermal cycling can be very large due to the large thermal expansion mismatch that exists between the silicon, on-chip dielectrics and metallization, leadframe, and the plastic molding compound used for chip encapsulation. Thus, to accelerate thermal cycling failure mechanism, the packaged chip may be accelerated by using temperature cycling ranges outside the normal range of operation and determining the number-of-cycles-to-failure. Some commonly used temperature cycling ranges for ULSI devices include: −65°C/150°C, −40°C/140°C, and 0°C/125°C.

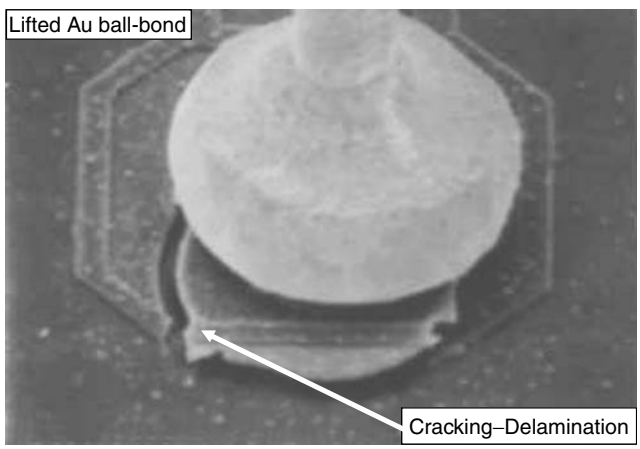


FIGURE 30.20 Lifted ball-bond has occurred due to thermomechanical stress during temperature cycling.

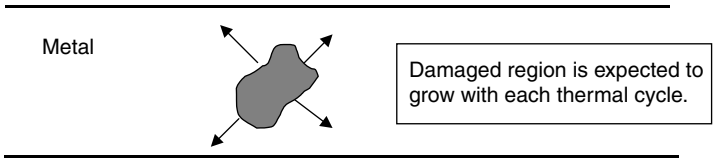


FIGURE 30.21 Thermomechanical stress during temperature cycling can cause plastic deformation (fatigue damage) for metals and their interfaces.

As for modeling, we assume that each thermal cycle generates plastic deformation which serves to damage the materials, as is illustrated in Figure 30.21. For ductile materials, low-cycle fatigue data are described rather well by the Coffin–Manson model [57–58]:

$$N_f = A_0(\Delta\varepsilon_p)^{-B} \tag{30.35}$$

where N_f, cycles to failure; $\Delta\varepsilon_p$, plastic strain range; and B, an empirically determined constant.

Low-cycle fatigue usually refers to stress conditions that only require a few hundred (or few thousand) cycles to produce failure. High-cycle fatigue usually refers to stress conditions which may require hundreds-of-thousands of cycles to produce failure.

During a temperature cycle, not all of the entire temperature range ΔT may be inducing plastic deformation. If a portion of this range, ΔT_0, is actually in the elastic range, then this should be subtracted and one can write a modified Coffin–Manson equation as [59]:

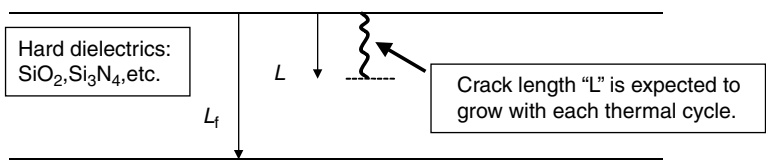
$$\Delta\varepsilon_p \propto (\Delta T - \Delta T_0)^{\beta} \tag{30.36}$$

Thus, for temperature cycling, the Coffin–Manson equation becomes:

$$N_f = A_0(\Delta T - \Delta T_0)^{-q} \tag{30.37}$$

where q is an empirically determined exponent. It is noted that if the elastic range (ΔT_0) is much smaller than the entire temperature cycle range (ΔT), then it may be dropped without significant error being introduced (the usual practice). As illustrated in Figure 30.22, fatigue can also occur in brittle materials due to crack propagation.

Normally there are three distinct phases to brittle material failure: a crack initiation phase (which usually exists at time zero), a crack growth phase (which tends to dominate the number of cycles to failure), and a catastrophic failure phase which is of very short duration. Since the crack- growth-phase is



Three phases to brittle-materials failure:
1. Crack initiation (usually exists at time zero)
2. Crack growth (dominates the cycles-to-failure)
3. Catastrophic failure (occurs very rapidly near end of cycles-to-failure)

FIGURE 30.22 Thermomechanical stress during temperature cycling can cause crack-propagation (fatigue damage) for brittle materials and their interfaces.

TABLE 30.2 Temperature Cycling Exponents

Material	q
Ductile metals (e.g., solder)	1–3
Hard-metal alloys/intermetallics (e.g., Al–Au)	3–5
Brittle materials (e.g., dielectrics: SiO_2, Si_3N_4)	6–9

of greatest duration (dominates the number of cycles to failure), the modeling effort is usually focused on this phase. Experimentally, we find that the crack-growth rate (increase in crack length per cycle) is dependent on the length of the existing crack and on the applied cyclical stress so one can write:

$$\frac{dL}{dN} = C(\sigma_a)^m L^n \tag{30.38}$$

where L is the crack length, N is the number of cycles, σ_a is the applied cyclical stress, m and n are empirically determined exponents. Separation of variables and integrating gives:

$$N_f = \left[\frac{1}{C}\right]\left[\int_{L_0}^{L_f} \frac{dL}{L^n}\right](\sigma_a)^{-m} = B_0(\sigma_a)^{-m} \tag{30.39}$$

Since the cyclical stress is assumed to be thermomechanical, $\sigma_a \propto \Delta T$, then cycles to failure becomes:

$$N_f = A_0(\Delta T)^{-q} \tag{30.40}$$

which is nearly identical to Equation 30.37. Thus, while the Coffin–Manson model was originally developed for ductile materials (metals), it can also be applied successfully to brittle materials with the appropriate selection of exponents as is summarized in Table 30.2.

In summary, temperature cycling failures for ULSI devices can be described reasonably well by the modified Coffin–Manson equation. The equation works rather well even for brittle material failures where failure is dominated by crack initiation and growth rather than simple plastic deformation (which were assumed in the development of the original Coffin–Manson equation) [59,60].

30.8 Time-Dependent Dielectric Breakdown

Due to the very high operating electric fields in the gate dielectric of CMOS devices, TDDB can be an important ULSI failure mechanism [63–98]. After usually a long period of degradation (trap creation/bond-breakage), as illustrated in Figure 30.23a, the dielectric eventually undergoes a catastrophic thermal run-a-way condition due to severe current flow. This localized severe joule heating can result in a conductive filament forming in the dielectric shorting the gate to the substrate (anode and cathode) in the metal oxide semiconductor (MOS) device (see Figure 30.23b). Two models have been widely used to describe the time-dependent failure mechanism in oxides. One model is field-driven (E-model) while the other is current-driven (1/E—model).

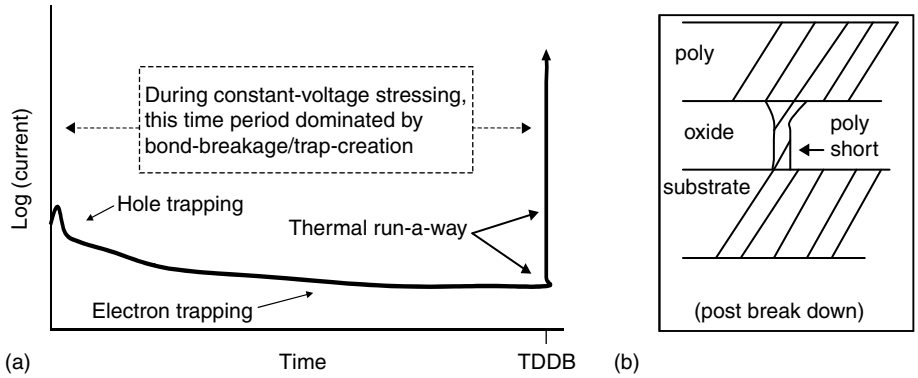


FIGURE 30.23 (a) Dielectric degradation occurs due to trap-generation/broken-bonds in the dielectric material and at the SiO_2/Si interface. (b) The trapping of the holes/electrons continues up to the point of catastrophic breakdown whereby the localized Joule heating produces a melt-filament shorting the gate and substrate.

30.8.1 E-Model (Field-Based Model)

In the thermochemical E-Model [63–75], the cause of low-field (<10 MV/cm) high temperature TDDB is due to field-enhanced thermal bond-breakage. In this model, the field serves to stretch polar molecular bonds thus making them weaker and more susceptible to breakage by standard Boltzmann (thermal) processes. Since the field reduces the activation energy required to break a bond [75], the degradation rate is expected to increase exponentially with field. Time-to-failure occurs when a localized density of broken bonds (or percolation sites) becomes sufficiently high to cause a conductive path to form from anode to cathode [76,77].

Time-to-failure, which is the inverse of degradation rate, decreases exponentially with field:

$$\mathrm{TF} = A_0 \exp\left(\frac{Q}{K_B T} - \gamma E_{ox}\right) \tag{30.41}$$

where γ, the field-acceleration parameter; E_{ox}, the electric field (usually expressed as MV/cm) in the oxide; Q, the activation energy (enthalpy of activation); and A_0, a process dependent coefficient.

[Note: many papers in the literature may use the base 10 (rather than the base e) when expressing the field dependence. One should be careful to note whether the base 10 or the natural base e is being used when a value is being reported for γ. For this reason, some authors for clarity will tend to write the field acceleration as *decades* per MV/cm to emphasize that the base 10 is being used or *Naperians* per MV/cm to emphasize that the natural base e is being used. However, not all authors will make this distinction so the reader must take caution. The conversion factor between the base 10 and base e is 2.3, i.e., $\gamma_{\text{base e}} = 2.3 \times \gamma_{\text{base 10}}$. In this work, the natural base e is assumed throughout.].

Many investigations have shown that γ is temperature dependent and can be described well by a simple $1/T$ dependence:

$$\gamma(T) = \frac{p_{\text{eff}}}{K_B T} \tag{30.42}$$

where p_{eff} is the effective dipole moment. p_{eff} is generally found to be in the 7–13 eÅ range [78] but can be much larger for higher dielectric constant materials [79]. The $1/T$ dependence for γ as is expressed in Equation 30.42 thus serves to drive observed/effective activation energy, obtained from Equation 30.13

and Equation 30.41, which tends to decrease linearly with the electric field,

$$Q_{\text{eff}} = Q - p_{\text{eff}} E_{\text{ox}} \tag{30.43}$$

where Q_{eff}, effective activation energy (eV); and Q, the enthalpy of activation for bond breakage in the absence of external field.

The observed γ, however, may not be necessarily temperature-dependent if several types of disturbed bonding states are present and participating in the dielectric degradation process under high-field and/or high-temperature TDDB testing [80]. Generally, however, for silica-based dielectrics with thicknesses >40 Å tested at 105°C, a $\gamma \sim 4.0$ cm/MV and a $Q \sim 1.5$ eV are generally observed during silica-based TDDB testing. This is true not only for silica-based gate dielectrics but is also true for silica-based low-k interconnect dielectrics [81]. For extrinsic defects, the effective oxide thickness (defective region of normal oxide) can be quite thin, i.e., the effective field can be very high for oxide defects, generally leading to apparently lower effective energies ($Q_{\text{eff}} \sim 0.3$ eV) being observed during burning.

30.8.2 1/E—Model (Current-Based Model)

In the 1/E model [82–85] for TDDB (even at low fields) is postulated to be due to current flow through the dielectric due to Fowler–Nordheim (F–N) conduction. Electrons, which are F–N injected from the cathode, may cause damage to the dielectric due to impact ionization as the electrons are accelerated through the dielectric. Also, when these accelerated electrons finally reach the anode, hot holes may be produced which can tunnel back into the dielectric causing damage (hot-hole anode-injection model). Since both the electrons from the cathode and the hot-holes from the anode are the result of F–N conduction, then the TF is expected to show an exponential dependence on the reciprocal of the electric field, 1/E:

$$\text{TF} = \tau_0(T) \exp\left[\frac{G(T)}{E_{\text{ox}}}\right] \tag{30.44}$$

where $\tau_0(T)$, a temperature dependent prefactor; G, the 1/E model field acceleration parameter.

The temperature dependence of G has been expressed as a $1/T$ power series expansion [84] given by,

$$G = G_0\left[1 + \left(\frac{\delta}{K_{\text{B}}}\right)\left(\frac{1}{T} - \frac{1}{300\ \text{K}}\right)\right] \tag{30.45}$$

where

$$\delta = \left(\frac{K_{\text{B}}}{G_0}\right)\left[\frac{\text{d}G}{\text{d}(1/T)}\right]_{300\ \text{K}} \tag{30.46}$$

and where the derivative is evaluated at 300 K. At room temperature [84], $G_0 \sim 350$ MV/cm and $\delta \sim 0.017$ eV. $\tau_0(T)$ is usually also represented as $1/T$ expansion,

$$\tau_0(T) = \tau_0 \exp\left[\left(\frac{-Q}{K_{\text{B}}}\right)\left(\frac{1}{T} - \frac{1}{300\ \text{K}}\right)\right] \tag{30.47}$$

where

$$\tau_0 \sim 1 \times 10^{-11}\ \text{s} \quad \text{and} \quad Q \sim 0.3\ \text{eV}.$$

30.8.3 Complementary Models (Inclusion of Both Field and Current into a Single Model)

There have been several attempts to include both field and current effects into a single TDDB model with some degree of success [86–88]. These modeling efforts permit both field-induced and current-induced dielectric degradation mechanisms occur simultaneously, in parallel fashion, during the TDDB testing. In Ref. 88 it was shown that current-induced hole-capture can serve to catalyze the bond breakage process, thus playing an important role in TDDB. Such hole capture can lead to very strong bonds being broken that would otherwise not be amenable to breakage by a field-only mechanism. Also, a hydrogen-release model has also been proposed [89,90] with a power-law model (with an exponent of $n \sim 40$) used for TF [91]. Both the hole-generation and the hydrogen-release models are expected to show a polarity dependence which is widely reported for hyper-thin (<4.0 nm) gate oxides [92]. Also, the adverse effects of hydrogen on TDDB have apparently been confirmed experimentally [93].

30.8.3.1 TDDB Summary

There has been great disagreement in the technical community as to the dominant degradation mechanism for low-field TDDB in SiO_2 thin films; i.e., is the major degradation mechanism related to: current or field? Certainly hole-capture and hydrogen-release are relevant mechanisms and must be folded into any TDDB discussion. While the E-model has been widely used and has been quite successful in describing low-field TDDB data for thick films >4.0 nm [63–75], however, for very thin oxides (<4.0 nm) the direct-tunneling current can be very high in these films and could mean that the degradation mechanism in hyperthin oxide films is more controlled by current than field.

Also, TDDB should not be considered just a gate oxide issue (see Figure 30.24). The issue of TDDB has also been raised for interconnects (metallization plus surrounding/supporting dielectrics) with the introduction of low-k dielectrics [94–98]. While low-k materials enable significant performance gains at the interconnect level in terms of circuit delay, they also possess substantially inferior electrical properties

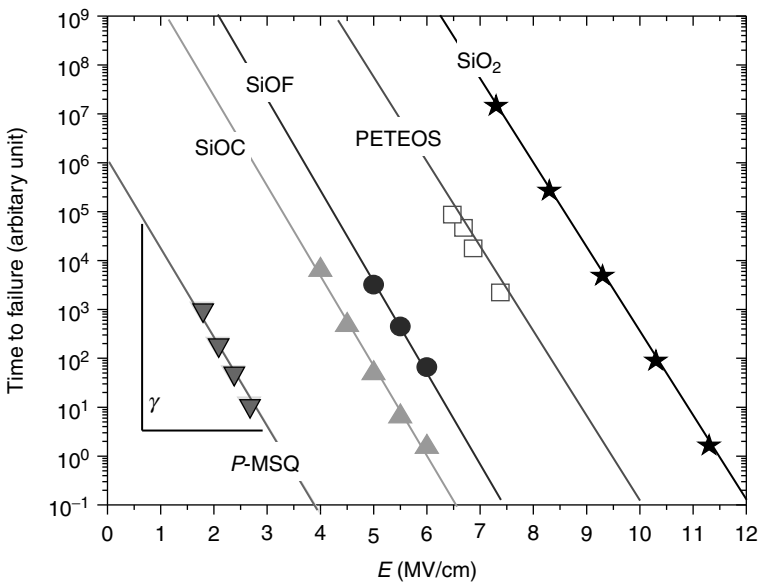


FIGURE 30.24 Lifetime data for various silica-based dielectrics at 105°C. The lower-k materials [MSQ($k=2.3$), SiOC($k=2.9$), and SiOF($k=3.6$)] generally have lower breakdown strength and time-to-failure. However, all of these silica based materials tested had a very similar acceleration parameter of $\gamma \sim 4$ cm/MV (or a peff ~ 13 eÅ).

to gate-oxide quality dielectric in terms of leakage and breakdown strength. Presently, inter-metal spacing between adjacent interconnect metal is approaching the physical dimensions of gate-oxides used a couple of decades ago. Hence, discussion of E- and 1/E-model physics is pertinent to low-*k* dielectrics as well. Oxide-based low-*k* dielectrics have been shown to have inferior breakdown strength and significantly wider failure distributions under constant voltage stress and are attributed to the presence of preexisting defects in the low-*k* dielectrics that scale roughly with the degree of porosity present within the low-*k* [95]. Yet, these low-*k* TDDB failures still appear to follow the same degradation physics found in gate oxides; i.e., similar field acceleration parameter $\gamma \sim 4$ cm/MV at 105°C (giving an effective dipole moment $p_{\text{eff}} \sim 13$ eÅ). A "pore" in this context is identified as a localized region of low-polarizability with existing weak bonds and charge traps in the material. Percolation modeling and the assumption of preexisting electrically active defects that scale with the degree of porosity naturally explains both the degraded breakdown strength and wider failure distributions of low-*k* dielectrics. Thus, careful assessments of the reliability margins present when using advanced low-*k* materials are a necessary part of successful process and integration of these new materials. Also, the contact to gate-edge spacing is presently only a few hundred Angstroms, similar to gate oxide thickness just a couple of decades ago. Thus, gate to contact TDDB must also be considered.

30.8.4 Mobile-Ions/Surface Inversion

Alkaline-metal elements such as Li, Na, and K can sometimes be found in the semiconductor processing materials. In SiO_2, these ions are very mobile under the presence of modest electric fields (~ 0.5 MV/cm) and temperatures (100°C). An accumulation (see Figure 30.25) of the drifted ions at the Si/SiO_2 interface can cause surface inversion and lead to increased leakage and device failure [99–103].

Sodium and potassium are the usual mobile-ion suspects, simply because of their high mobility and their relative abundance in many materials. Under bias, they can drift from the poly anode to the silicon substrate (cathode). A buildup of positive ions at the Si/SiO_2 interface can invert the surface and severely degrade the oxide isolation. Ionic drift in SiO_2 gate dielectric can also cause premature TDDB. In the case of erasable programmable read only memory (EPROMs), mobile-ion accumulation around the negatively-charged floating poly can lead to data retention fails [103].

Devices showing such inversion induced leakage failures can recover during an unbiased high temperature bake. The bake causes a redistribution of the mobile-ions away from the accumulated Si/SiO_2 interface (or floating poly in the case of an EPROM-like device).

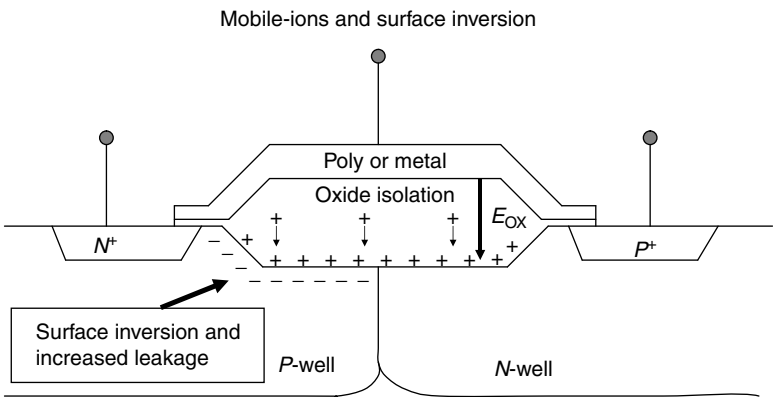


FIGURE 30.25 Mobile-ions can drift in the field oxide isolation due to the electric field and eventually cause surface inversion of the P-well. This surface inversion can cause a leakage path from *N*-well to the adjacent N^+ moat (in *P*-well).

Since the mobile-ion flux is impacted both electric field and temperature, the TF is usually described by:

$$\mathrm{TF} = A_0 J_{\mathrm{ion}}^{-1} \exp\left(\frac{Q}{K_{\mathrm{B}} T}\right) \tag{30.48}$$

and,

$$J_{\mathrm{ion}} = \left\langle \left(\frac{D_0}{K_{\mathrm{B}} T}\right)(eE) - D_0 \frac{\partial \rho(x,t)}{\partial x} \right\rangle \tag{30.49}$$

where J_{ion} is the time-average flow of ions. The first term on the right hand side of Equation 30.49 is the drift component with E, the externally applied electric field; D_0, diffusion coefficient; and ρ, density of mobile ions. The second term is the back-diffusion component and the brackets $\langle\,\rangle$ represent the time-averaged value of the time-dependent quantities enclosed. Note that if the field is turned off and the device is baked (an unbiased bake), the J_{ion} can change direction (will now be dominated by back diffusion) and the surface-inversion can be overcome and the device can recover. This is referred to as bake-recovery fail. The activation energy Q depends upon the ULSI medium through which the ion must diffuse, and for Na it ranges from 0.75 to 1.8 eV, with 1.0 being typical. [Note: Stuart, Ref. 99, finds 0.75 eV for Na (theory and experiment) and $\sim$1.1 eV or more for everything else.].

Mobile Cu ions under electrical bias are also a concern in the backend, where loss of barrier integrity or the presence of Cu-related corrosion defects will lead to substantially degraded backend dielectric reliability performance [94–96]. Since interfaces are prevalent within Cu interconnect geometries, fast migration pathways are available for relatively rapid migration. Since such defects are difficult to observe under at-use conditions, their statistical presence must be determined using accelerated test conditions and rapid tests such as ramped breakdown [97,98]. Usually, Cu ion drift under electric field is an issue more for interconnect TDDB than for surface inversion.

In summary, for mobile-ions the activation energy for ion diffusion depends on: the diffusing species, medium through which the mobile ions must diffuse, and the concentration of the ions. If the mobile-ion concentration is low, and if deep interfacial traps exist, then one may see a deep interfacial-trap dominated activation energy of $\sim$1.8 eV for Na+ diffusion through SiO_2. However, if the concentration of Na+ is high such that all interfacial traps can be filled and highly mobile Na+ ions are left over, then one might see a lower activation energy of $Q \sim 0.75$ eV for simple Na+ drift through SiO_2 (simple interstitial-like diffusion through undoped oxides).

30.8.5 Channel Hot-Carrier Injection

Channel hot-carrier injection (HCI) describes the phenomena by which electrons (or holes) can gain sufficient kinetic energy that they can be injected over the 3.1 eV barrier that exists at the Si/SiO_2 interface for electrons. Channel electrons, as they are accelerated along the channel of a metal oxide semiconductor field effect transistor (MOSFET) device (see Figure 30.26), can acquire the needed energy, especially for those "lucky electrons" located near the tail of the Boltzmann distribution. The lucky or hot carriers moving along the channel can be injected into the gate oxide as a result of impact ionization near the drain end of the MOSFET device where the electric field is the greatest. Channel hot-carrier injection can serve to produce damage at the interface (interface state generation). Interface-state generation and charge trapping by this HCI mechanism can result in transistor parameter degradation. This is an important degradation mechanism, especially for all advanced technologies where the channel electric fields which accelerate the carriers have increased faster than the reductions in operating voltage. Thus, HCI can be an important ULSI failure mechanism [104–111]. Generally, the degradation induced by HCI can be described by:

$$\Delta p = B_0 t^n \tag{30.50}$$

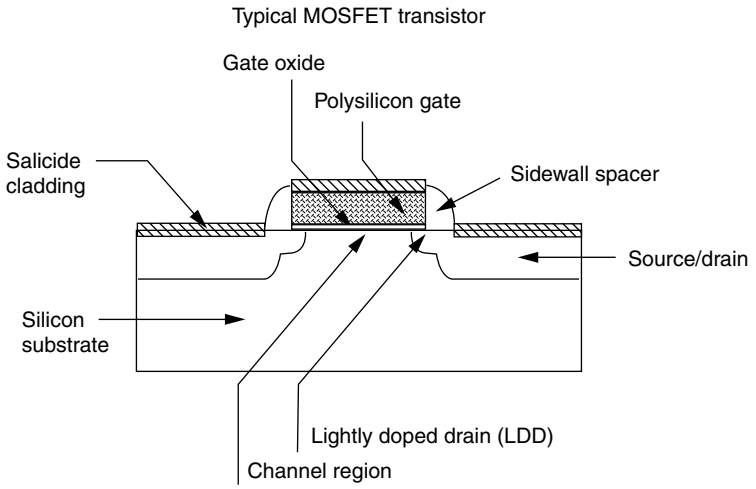


FIGURE 30.26 Carriers traveling along the MOSFET channel are accelerated from source to drain. Impact ionization can occur at the drain end (where the electric field is greatest) causing some of the carriers to be redirected toward the gate oxide and where they can produce interface-state generation resulting in device degradation (changes in critically important device parameters, e.g., V_t, gm, I_{drive}, I_{off}, etc.).

where p, the parameter of interest (V_t, gm, Id_{sat}, etc.); t, the time; B_0, a material dependent parameter; and n, a function of stressing voltage V, temperature and effective transistor channel length L_{eff}, i.e., $n = n(V, L_{eff}, T)$.

For N-channel devices, when the hot electron undergoes impact ionization near the drain end of the device, holes are produced which can be collected as substrate current I_{sub}. The peak I_{sub} current thus becomes an easy-to-measure indicator of the channel hot carrier stress that is being applied to the device during stress. The TF expression generally accepted for N-channel transistors is:

$$TF = A_0 I_{sub}^{-n} \exp\left(\frac{Q}{K_B T}\right) \tag{30.51}$$

where I_{sub}, peak substrate current during stressing; $n \sim 3$, and Q, -0.1 to -0.2 eV.

For P-channel devices, sometimes the gate current I_{gate} is the better monitor for the stress on the device. Thus, for P-channel devices the TF equation for CHI may be written:

$$TF = A_0 I_{gate}^{-n} \exp\left(\frac{Q}{K_B T}\right) \tag{30.52}$$

where $I_{gate=}$peak gate current during stressing; $n = 2$–4; $Q = -0.1$ to -0.2 eV.

In summary, HCI-induced transistor degradation seems to be satisfactorily modeled by peak substrate-current I_{sub} for the N-channels and peak gate current I_{gate} for the P-channels, at least for transistors at > 0.25 μm. While the drive-current for the N-channel device tends to reduce after HCI stressing (i.e., HCI stressing tends to produce charge trapping such that is serves to increase the effective channel length for the N-channel device), the P-channel drive current tends to increase after HCI stress (i.e., HCI stressing tends to produce charge trapping such that the degradation serves to increase and shorten the effective channel length for the P-channel devices) and the off-state leakage can increase dramatically [111].

While HCI-induced transistor degradation measurement and modeling seems to be very accurate, the extrapolation from transistor degradation to circuit-level degradation is very uncertain and should be

approached with caution. Also, there is growing evidence that HCI physics may be starting to change below 0.25 μm channel lengths and worst-case stressing conditions may have to be changed.

30.9 Summary and a Look into the Future

The dominance of CMOS technology over the last two decades has permitted the industry to focus its efforts primarily on scaling and this has led to remarkable success in terms of performance, functionality, and cost per function. During this same time period, tremendous improvements in chip reliability have also been accomplished through extensive reliability-physics understanding and proactive reliability-engineering efforts.

With 20+ years of continuous CMOS scaling (which has been required for improved circuit density and performance), the existing CMOS materials have now been pushed to their physical and reliability limits. In terms of physical property limitations: aluminum has become too resistive, and the industry conversion to copper is well underway. In addition, the supporting and surrounding interconnect dielectrics have dielectric constants k which are too high, and the need will become greater for lower-k dielectrics to help minimize the RC-type interconnect delays. However, generally these lower-k materials have lower electrical breakdown-strength and lower mechanical-strength, so interconnect reliability issues will have to be studied closely. Also, CMOS scaling of the standard gate oxide will likely end due to leakage and/or breakdown issues. This will necessitate the investigation of new high-k gate dielectrics and perhaps metal gates (rather than the standard doped-polysilicon gates). Again, the reliability issues associated with high-k gate-dielectrics and metal gate-electrodes will have to be studied closely.

In summary, reliability physics has played a very key role in the success of the semiconductor industry. Furthermore, it is projected that reliability physics will play an even more important role as we try to replace the core CMOS materials which have served the industry so well for so long: copper must replace aluminum, low-k dielectrics must replace standard SiO_2 interconnect dielectrics, high-k gate-dielectric must replace SiO_2 gate-oxide, and metal gate-electrodes will likely have to replace conventional doped-polysilicon. A clear understanding of the reliability physics for these new materials will be critically important for their successful and timely introduction into the semiconductor industry.

References

1. Endicot, H., and T. Walsh. "Accelerated Testing of Component Parts." *Annu. Symp. Reliab.* (1966): 571.
2. Bora, J., and A. Babar. "Simplification of Base Failure Rate Models." *Microelectron. Reliab.* (1980): 535.
3. McPherson, J. W. "Stress Dependent Activation Energy." In *IEEE International Reliability Physics Symposium Proceedings*, 12, 1986.
4. LaCombe, D., and E. Parks. "The Distribution of Electromigration Failures." In *IEEE International Reliability Physics Symposium Proceedings*, 1, 1986.
5. Gall, M., et al. "Scaling Effect on Electromigration in On-Chip Cu Wiring." In *IEEE International Interconnect Technology Conference Proceedings*, 270, 1999.
6. Degraeve, R. "Oxide Reliability." In *IEEE International Reliability Physics Symposium Tutorials*, 7, 1997.
7. Hunington, H., and A. Grone. "Current Induced Marker Motion in Gold Wires." *J. Phys. Chem. Solids* 20 (1961): 76.
8. Black, J. "A Brief Survey of Some Recent Electromigration Results." *IEEE Trans. Elect. Dev.* ED-16 (1969): 338.
9. d-Heurle, F., and P. Ho. "Electromigration in Thin Films." In *Thin Films: Interdiffusion and Reactions*, 243. Wiley: New York, 1978.

10. Blech, I., and H. Sello. "The Failure of Thin Aluminum Current-Carrying Strips on Oxidized Silicon." In *Physics of Failures in Electronics*, Vol. 5, USAF-RADC Series, 496, 1966.
11. Martin, C., and J. McPherson. "Via Electromigration Performance of Ti/W/Al–Cu(2%) Multi-Layered Metallization." In *VLSI Multilevel Interconnect Conference Proceedings*, 168, 1989.
12. Filippi, R., G. Biery, and R. Wachnik. "The Electromigration Short-Length Effect in Ti–AlCu–Ti Metallization with Tungsten Plugs." *J. Appl. Phys.* 78 (1995): 3756.
13. Oates, A. "Electromigration Failure Distrubution of Contacts and Vias as a Function of Stress Conditions in Submicron IC Metallizations." In *IEEE International Reliability Physics Symposium Proceedings*, 164, 1996.
14. Vaidya, S., et al. "Electromigration Induced Shallow Junction Leakage with Al/Poly-Si Metallization." *J. Electrochem. Soc.* 130 (1983): 496.
15. Vaidya, S., et al. "Shallow Junction Cobalt Silicide Contacts with Enhanced Electromigration Resistance." *J. Appl. Phys.* 55 (1984): 3514.
16. Steenwyk, S., and E. Kankowski. "Electromigration in Aluminum to Ta-Silicide Contacts." In *IEEE International Reliability Physics Symposium Proceedings*, 30, 1986.
17. Ondrusek, J., C. Dunn, and J. McPherson. "Kinetics of Contact Wearout for Silicided($TiSi_2$) and Nonsilicided Contacts." In *IEEE International Reliability Physics Symposium Proceedings*, 154, 1987.
18. Maiz, J. "Characterization of Electromigration under Bidirectional and Pulsed Unidirectional Currents." In *IEEE International Reliability Physics Symposium Proceedings*, 220, 1989.
19. Ting, L., J. May, W. Hunter, and J. McPherson. "ac Electromigration Characterization and Modeling of Multilayered Interconnects." In *IEEE International Reliability Physics Symposium Proceedings*, 311, 1993.
20. Graas, C., H. Le, J. McPherson, and R. Havemann. "Electromigration Reliability Improvements of W-Plug Vias by Titanium Layering." In *IEEE International Reliability Physics Symposium Proceedings*, 173, 1994.
21. McPherson, J., H. Le, and C. Graas. "Reliability Challenges for Deep Submicron Interconnects." *Microelectron. Reliab.* 37 (1997): 1469.
22. Hu, C., et al. "Scaling Effect on Electromigration in On-Chip Cu Wiring." In *IEEE International Interconnect Conference*, 267, 1999.
23. Ogawa, E. T., K.-D. Lee, V. A. Blaschke, and P. S. Ho. "Electromigration Reliability Issues in Dual-Damascene Cu Interconnections." *IEEE Trans. Reliab.* 51, no. 4 (2002): 403.
24. Hussein, M. A., and J. He. *IEEE Trans. Semicond. Manuf.* 18, no. 01 (2005): 69.
25. Ogawa, E. T., et al. "Statistics of Electromigration Early Failures in Cu/Oxide Dual-Damascene Interconnects." In *39th Annual IEEE International Reliability Physics Symposium Proceedings*, 341, 2001.
26. Hau-Riege, C. S. "An Introduction to Cu Electromigration." *Microelectron. Reliab.* 44 (2004): 195.
27. Lane, W. M., E. Liniger, and J. R. Lloyd. "Relationship between Interfacial Adhesion and Electromigration in Cu Metallization." *J. Appl. Phys.* 93, no. 3 (2003): 1417.
28. Hu, C.-K., et al. "Effects of Overlayers on Electromigration Reliability Improvement for Cu/Low *K* Interconnects." In *42th Annual IEEE International Reliability Physics Symposium Proceedings*, 222, 2004.
29. Lee, K-D. , and P. S. Ho. *IEEE Trans. Dev. Mater. Reliab.* 4, no. 2 (2004): 237.
30. Michael, N. L., C.-U. Kim, P. Gillespie, and R. Augur. *Appl. Phys. Lett.* 83, no. 10 (1959): 2003.
31. Park, Y.-J., Park, K.-D. Lee, and W. R. Hunter. In *43th Annual IEEE International Reliability Physics Symposium Proceedings*, 18, 2005.
32. Hau-Riege, S. P. "Probabilistic Immortality of Cu Damascene Interconnects." *J. Appl. Phys.* 91, no. 4 (2002): 2014.
33. Hau-Riege, C., A. P. Marathe, and V. Pham. In *41st Annual IEEE International Reliability Physics Symposium Proceedings*, 173, 2003.
34. Lee, K.-D., X. Lu, E. T. Ogawa, H. Matsuhashi, and P. S. Ho. "Electromigration Study of Cu/low *k* Dual-Damascene Interconnects." In *40th Annual IEEE International Reliability Physics Symposium Proceedings*, 322, 2002.

35. Schnable, A., and R. Keen. "Failure Mechanisms in Large-Scale Integrated Circuits." In *IEEE International Reliability Physics Symposium Proceedings*, 170, 1969.
36. Peck, D. "The Design and Evaluation of Reliable Plaastic-Encapsulated Semiconductor Devices." In *IEEE International Reliability Physics Symposium Proceedings*, 81, 1970.
37. Koelmans, H. "Metallization Corrosion in Silicon Devices by Moisture-Induced Electrolyis." In *IEEE International Reliability Physics Symposium Proceedings*, 168, 1974.
38. Flood, J. "Reliability Aspects of Plastic Encapsulated Integrated Circuits." In *IEEE International Reliability Physics Symposium Proceedings*, 95, 1972.
39. Paulson, W., and R. Kirk. "The Effects of Phosphorus-Doped Passivation Glass on the Corrosion of Aluminum." In *IEEE International Reliability Physics Symposium Proceedings*, 172, 1972.
40. Lawrence, J., and J. McPherson. "Corrosion Susceptibility of Al–Cu and Al–Cu–Si Films." *J. Electrochem. Soc.* 137 (1990): 3879.
41. Gunn, J., R. Camenga, and S. Malik. "Rapid Assessment of the Humidity Dependence of IC Failure Modes by Use of Hast." In *IEEE International Reliability Physics Symposium Proceedings*, 66, 1983.
42. Peck, D. "A Comprehensive Model for Humidity Testig Correlation." In *IEEE International Reliability Physics Symposium Proceedings*, 44, 1986.
43. Dunn, C., and J. McPherson. "Recent Observations on VLSI Bond Pad Corrosion Kinetics." *J. Electrochem. Soc.* 135 (1988): 661.
44. McPherson, J., G. Bishel, and J. Ondrusek. In *Proceedings of Third International Symposium on Corrosion and Reliability of Electronic Materials and Devices*, Vol. 94-29, 270. Electrochemical Society, 1994.
45. Klema, J., R. Pyle, and E. Domangue. "Reliability Implications of Nitrogen Contaminated during Deposition of Sputtered Aluminum/Silicon Metal Films." In *IEEE International Reliability Physics Symposium Proceedings*, 1, 1984.
46. Yue, J., W. Fusten, and R. Taylor. "Stress Induced Voids in Aluminum Interconnects during IC Processing." In *IEEE International Reliability Physics Symposium Proceedings*, 126, 1985.
47. McPherson, J., and C. Dunn. "A Model for Stress-Induced Metal Notching and Voiding in VLSI Al–Si Metallization." *J. Vac. Sci. Technol. B*, 5 (1987): 1321.
48. Groothuis, S., and W. Schroen. "Stress Related Failures Causing Open Metallization." In *IEEE International Reliability Physics Symposium Proceedings*, 1, 1987.
49. Yue, J. T. "Reliability." In *ULSI Technology*, 674. New York: McGraw-Hill, 1996.
50. McPherson, J. "Accelerated Testing." In *Electronic Materials Handbook, Volume 1 Packaging*, 887. Materials Park, OH: ASM International Publishing, 1989.
51. Paik, J.-M., J.-K. Jung, and Y.-C. Joo. "The Dielectric Material Dependence of Stress and Stress Relaxation on the Mechanism of Stress-Voiding of Cu Interconnects." In *43th Annual IEEE International Reliability Physics Symposium Proceedings*, 195, 2005.
52. Harper, J. M. E., C. Cabral Jr., P. C. Andricacos, L. Gignac, I. C. Noyan, K. P. Rodbell, and C. K. Hu. "Mechanisms for Microstructure Evolution in Electroplated Copper Thin Films near Room Temperature." *J. Appl. Phys.* 86, no. 5 (1999): 2516.
53. Edelstein, D., J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, P. Roper, et al. "Full Copper Wiring in a Sub-0.25 μm CMOS ULSI Technology." In *IEEE International Electron Devices Meeting Technical Digest*, 773, 1997.
54. Ogawa, E. T., J. W. McPherson, J. A. Rosal, K. J. Dickerson, T.-C. Chiu, L. Y. Tsung, M. K. Jain, T. D. Bonifield, J. C. Ondnsek, and W. R. McKee. "Stress-Induced Voiding under Vias Connected to Wide Cu Metal Leads." In *40th Annual IEEE International Reliability Physics Symposium Proceedings*, 312, 2002.
55. Yoshida, K., T. Fujimaki, K. Miyamoto, T. Honma, H. Kaneko, H. Nakazawa, and M. Morita. "Stress-Induced Voiding Phenomena for an Actual CMOS LSI Interconnects." In *IEEE International Electron Devices Meeting Technical Digest*, 753, 2002.
56. von Glasow, A., A. H. Fischer, M. Hierlemann, S. Penka, and F. Ungar. "Geometrical Aspects of Stress-Induced Voiding in Copper Interconnects." In *Advanced Metallization Conference Proceedings* (*AMC*), 161, 2002.
57. Coffin, L., Jr. *Met. Eng. Q,* 3 (1963): 15.

58. Manson, S. *Thermal Stress and Low-Cycle Fatigue*. New York: McGraw-Hill Book Co., 1966.
59. Dunn, C. F., and J. W. McPherson. "Temperature Cycling Acceleration Factors in VLSI Applications." In *IEEE International Reliability Physics Symposium Proceedings*, 252, 1990.
60. Blish, R. C. II. "Temperature Cycling and Thermal Shock Failure Rate Modeling." In *IEEE International Reliability Physics Symposium Proceedings*, 110, 1997.
61. Caruso, H., and A. Dasgupta. "A Fundamental Overview of Accelerated-Testing Analytical Models." In *Proceedings of Annual Reliability and Maintainability Symposium*, 389, 1998.
62. Dieter, G. *Mechanical Metallurgy*, 467. New York: McGraw-Hill, 1976.
63. Anolick, E. S., and G. R. Nelson. "Low-Field Time-Dependent Dielectric Integrity." In *IEEE International Reliability Physics Symposium Proceedings*, 8, 1979.
64. Crook, D. "Method of Determining Reliability Screens for Time-Dependent Dielectric Breakdown." In *IEEE International Reliability Physics Symposium Proceedings*, 1, 1979.
65. Berman, A. "Time Zero Dielectric Reliability Test by a Ramp Method." In *IEEE International Reliability Physics Symposium Proceedings*, 204, 1991.
66. McPherson, J., and D. Baglee. "Acceleration Factors for Thin Gate Oxide Stressing." In *IEEE International Reliability Physics Symposium Proceedings*, 1, 1985.
67. McPherson, J., and D. Baglee. *J. Electrochem. Soc.* 132 (1903): 1985.
68. Swartz, G. "Gate Oxide Integrity of NMOS Transistor Arrays." *IEEE Trans. Elect. Dev.* ED-33 (1986): 1826.
69. Boyko, K., and D. Gerlach. "Time Dependent Dielectric Breakdown of 210A Oxides." In *IEEE International Reliability Physics Symposium Proceedings*, 1, 1989.
70. Suehle, J., et al. "Field and Temperature Acceleration of Time-Dependent Dielectric Breakdown in Intrinsic Thin SiO_2." In *IEEE International Reliability Physics Symposium Proceedings*, 120, 1994.
71. Charparala, P., et al. "Electric Field Dependent Dielectric Breakdown of Intrinsic SiO_2 Films under Dynamic Stress." In *IEEE International Reliability Physics Symposium Proceedings*, 61, 1996.
72. Suehle, J., and P. Chaparala. "Low Electric Field Breakdown of Thin SiO_2 Films under Static and Dynamic Stress." *IEEE Trans. Elect. Dev.* 44, no. 5 (1997): 801.
73. Kimura, M. "Oxide Breakdown Mechanism and Quantum Physical Chemistry for Time-Dependent Dielectric Breakdown." In *IEEE International Reliability Physics Symposium Proceedings*, 190, 1997.
74. McPherson, J., et al. "Field-Enhanced Si–Si Bond-Breakage Mechanism for Time-Dependent Dielectric Breakdown in Thin-SiO_2 Dielectrics." *Appl. Phys. Lett.* 71 (1997): 1101.
75. McPherson, J., and H. Mogul. "Underlying Physics of the Thermochemical E-Model in Describing Low-Field Time-Dependent Dielectric Breakdown in SiO_2 Thin Films." *J. Appl. Phys.* 84 (1998): 1513.
76. Degraeve, R., et al. "New Insights in the Relation between Electron Trap Generation and the Statistical Properties of Oxide Breakdown." *IEEE Trans. Elect. Dev.* 45 (1998): 904.
77. Sune, J., D. Jimenez, and E. Miranda. "Breakdown Modes and Breakdown Statistics of Ultrathin SiO_2 Gate Oxides." *J. High Speed Electron. Syst.* 11 (2001): 789.
78. McPherson, J. "Determination of the Nature of Molecular Bonding in Silica from Time-Dependent Dielectric Breakdown Data." *J. Appl. Phys.* 95 (2004): 8101.
79. McPherson, J. "Trends in the Ultimate Breakdown Strength of High Dielectric-Constant Materials." *IEEE Trans. Elect. Dev.* 50 (2003): 1771.
80. McPherson, J., and H. Mogul. "Disturbed Bonding States in SiO_2 Thin-Films and Their Impact on Time-Dependent Dielectric Breakdown." In *IEEE International Reliability Physics Symposium Proceedings*, 47, 1998.
81. Ogawa, E., J. Kim, and J. McPherson. "Leakage, Breakdown, and TDDB Characteristics of Porous Low-*k* Silica-Based Interconnect Dielectrics." In *IEEE International Reliability Physics Symposium Proceedings*, 166, 2003.
82. Chen, I., S. Holland, and C. Hu. "A Quantitative Physical Model for Time-Dependent Breakdown." In *IEEE International Reliability Physics Symposium Proceedings*, 24, 1985.
83. Lee, J., I. Chen, and C. Hu. "Statistical Modeling of Silicon Dioxide Reliability." In *IEEE International Reliability Physics Symposium Proceedings*, 131, 1988.

84. Moazzami, R., J. Lee, and C. Hu. "Temperature Acceleration of Time-Dependent Dielectric Breakdown." *IEEE Trans. Elect. Dev.* 36 (1989): 2462.
85. Schuegraph, K., and C. Hu. "Hole Injection Oxide Breakdown Model for Very Low Voltage Lifetime Extrapolations." In *IEEE International Reliability Physics Symposium Proceedings*, 7, 1993.
86. Hu, C., and Q. Lu. "A Unified Gate Oxide Reliability Model." In *IEEE International Reliability Physics Symposium Proceedings*, 47, 1999.
87. Cheung, K. P. "A Physics-Based, Unified Gate-Oxide Breakdown Model." In *Technical Digest of Papers International Electron Devices Meeting*, 719, 1999.
88. McPherson, J., R. Khamankar, and A. Shanware. "Complementary Model for Intrinsic Time-Dependent Dielectric Breakdown in SiO_2 Dielectrics." *J. Appl. Phys.* 88 (2000): 5351.
89. DiMaria, D., and J. Stasiak. "Trap Creation in Silicon Dioxide Produced by Hot Electrons." *J. Appl. Phys.* 65 (1989): 2342.
90. DiMaria, D., E. Cartier, and D. Arnold. "Impact Ionization, Trap Creation, Degradation, and Breakdown in Silicon Dioxide Films on Silicon." *J. Appl. Phys.* 73 (1993): 3367: Stathis, J., and D. DiMaria. *Technical Digest of Papers International Electron Devices Meeting*, 167, 1998.
91. Wu, E., et al. "Experimental Evidence of TBD Power-Law for Voltage Dependence of Oxide Breakdown in Ultrathin Gate Oxides." *IEEE Trans. Elect. Dev.* 49 (2002): 2244.
92. Wu, E., et al. "Polarity-Dependent Oxide Breakdown of NFET Devices for Ultra-Thin Gate Oxide." In *IEEE International Reliability Physics Symposium*, 60, 2002.
93. Pompl, T., et al. "Change in Acceleration Behavior of Time-Dependent Dielectric Breakdown by the BEOL Process: Indications for Hydrogen Induced Transition in Dominant Degradation Mechanism." In *IEEE International Reliability Physics Symposium*, 388, 2005.
94. R. Tsu, J. W. McPherson, and W. R. McKee, "Leakage and Breakdown Reliability Issues Associated with Low-*k* Dielectrics in a Dual-Damascene Cu Process." In *38th Annual IEEE International Reliability Physics Symposium Proceedings*, 348, 2000.
95. Ogawa, E. T., J. Kim, G. S. Haase, H. C. Mogul, and J. W. McPherson. "Leakage, Breakdown, and TDDB Characteristics of Porous Low-*k* Silica-Based Interconnect Dielectrics." In *41st Annual IEEE International Reliability Physics Symposium Proceedings*, 166, 2003.
96. Noguchi, J., N. Miura, M. Kubo, T. Tamaru, H. Yamaguchi, N. Hamada, K. Makabe, R. Tsuneda, and K. Takeda. "Cu-Ion-Migration Phenomena and Its Influence on TDDB Lifetime in Cu Metallization." In *41st Annual IEEE International Reliability Physics Symposium Proceedings*, 287, 2003.
97. Eissa, M. M., D. A. Ramappa, E. Ogawa, N. Doke, E. M. Zielinski, C. L. Borst, G. Shinn, and A. J. McKerrow. "Post-Copper CMP Cleans Challenges for 90 nm Technology." In *Advanced Metallization Conference Proceedings* (*AMC*), 559, 2004.
98. Haase, G. H., E. T. Ogawa, and J. W. McPherson. "Breakdown Characteristics of Interconnect Dielectrics." In *43th Annual IEEE International Reliability Physics Symposium Proceedings*, 466, 2005.
99. Stuart, D. A. "Calculations of Activation Energy of Ionic Conductivity in Silica Glass by Classical Methods." *J. Am. Ceramic Soc.* (1954): 573.
100. Snow, E. H., A. S. Grove, B. E. Deal, and C. T. Sah. "Ion Transport Phenomenon in Insulating Films." *J. Appl. Phys.* 36 (1965): 1664.
101. Snow, E. H., and B. E. Deal. "Polarization Phenomena and Other Properties of Phosphosilicate Glass Films on Silicon." *J. Electrochem. Soc.* 113 (1966): 263.
102. Schnable, A. "Failure Mechanisms in Microelectronic Devices." *Microelectron. Reliab.* (1988).
103. Hefley, P. L. and J. McPherson. "The Impact of an External Sodium Diffusion Source on the Reliability of MOS Circuitry." In *IEEE International Reliability Physics Symposium Proceedings*, 167, 1988.
104. Aur, S., A. Chatterjee, and T. Polgreen. "Hot Electron Reliability and ESD Latent Damage." In *IEEE International Reliability Physics Symposium Proceedings*, 15, 1988.
105. Takeda, E., R. Izawa, K. Umeda, and R. Nagai. "ac Hot-Carrier Effects in Scaled MOS Devices." In *IEEE International Reliability Physics Symposium Proceedings*, 118, 1991.

106. Snyder, E., D. Cambell, S. Swanson, and D. Pierce. "Novel Self-Stressing Test Structures for Realistic High-Frequency Reliability Characterization." In *IEEE International Reliability Physics Symposium Proceedings*, 57, 1993.
107. Wang-Ratkovic, J., et al. "New Understanding of LDD CMOS Hot-Carrier Degradation and Device Lifetime at Cryogenic Temperatures." In *IEEE International Reliability Physics Symposium Proceedings*, 312, 1997.
108. LaRosa, G., et al. "NBTI Channel Hot Carrier Efffects in PMOSFETS in Advanced CMOS Technologies." In *IEEE International Reliability Physics Symposium Proceedings*, 282, 1997.
109. Yue, J. T. "Reliability." In *ULSI Technology*, New York: McGraw-Hill, 657, 1996.
110. Fang, P., J. T. Yue, and D. Wollesen, "A Method to Project Hot-Carrier Induced Punch Through Voltage Reduction for Deep Submicron LDD PMOS FETs at Room and Elevated Temperatures." In *IEEE International Reliability Physics Symposium Proceedings*, 131, 1982.
111. Ong, T., P. Ko, and C. Hu. "Hot-Carrier Current Modeling and Device Degradation in Surface Channel PMOSFET." *IEEE Trans. Elect. Dev.* ED-37 (1990): 1658.

31

Effects of Terrestrial Radiation on Integrated Circuits

31.1 Background—Motivation and Terminology **31**-1

31.2 The Terrestrial Radiation Environment **31**-2
Ions in Matter • High-Energy Cosmic Ray Neutrons • Alpha Particles from the Radioactive Decay of Impurities • Low-Energy Cosmic Ray Neutrons and ^{10}B

31.3 Radiation Effects on Semiconductor Devices **31**-10

31.4 Technology Scaling Trends **31**-13
Memory SER Sensitivity • Sequential/Combinational Logic SER Sensitivity

31.5 Commercial Mitigation Techniques **31**-16
Source Mitigation Techniques • Process Technology Mitigation Techniques • Design Mitigation Techniques • System-Level Redundancy Techniques

31.6 Summary **31**-21

References **31**-21

Robert Baumann
Texas Instruments, Inc.

31.1 Background—Motivation and Terminology

As the dimensions and operating voltages of microelectronic are aggressively reduced to satisfy the consumer's insatiable demand for higher density, increased functionality, and reduced power consumption, their sensitivity to radiation has increased dramatically. Radiation effects in semiconductor devices are responsible for a plethora of reliability issues that vary in magnitude from single-bit data disruptions to events that lead to complete device destruction [1,2]. Of primary concern for commercial applications are the single-event effects (SEE) produced by several types of energetic particles present in the terrestrial environment.

Soft errors are caused by one of several possible SEE that can induce enough of a charge disturbance to reverse or flip the data state of one or more memory cells, registers, latches, or flip-flops. The error is characterized as "soft" because the circuit is not permanently damaged by the radiation; thus if new data is written to the bit, the device will store it correctly. In contrast, a "hard" error is manifested when the device is physically damaged such that improper operation occurs, data is lost, and the damaged state is permanent. Another intermediate failure mode known as an "intermittent" failure can sometimes be mistaken for a soft error, but is more closely related to a hard failure since it usually reoccurs from time to time in a marginal device or in one with latent damage that, under certain operating conditions, ceases to function properly.

The rate at which soft errors occur is called the soft error rate (SER). The unit of measure commonly used with SER and other reliability mechanisms is the failure-in-time (FIT), equivalent to an average failure rate of 1 error in 10^9 h. Soft errors have become a huge concern in advanced computer chips because, uncorrected, they produce a failure rate that is higher than all the other reliability mechanisms combined! For example, a typical failure rate for a hard reliability mechanism (such as gate oxide breakdown, transistor hot-carrier effects, metal lead electromigration, etc.) in a qualified product is somewhere between 1 and 50 FIT. There are several critical reliability mechanisms that can degrade integrated circuit performance, but in general the overall failure rate is typically 10–100 FIT for a qualified commercial product. In stark contrast, without mitigation, the SER can easily exceed 50,000 FIT. While 50,000 FIT represents less than one failure-per-year (assuming 24 h per day operation) for some high reliability applications, critical life-support systems or large systems with thousands of chips, this failure rate is unacceptable. Left unchallenged, soft errors have the potential for inducing the highest failure rate of all other reliability mechanisms combined. Many of today's advanced microprocessors and digital signal processors incorporate one or more soft error mitigation schemes to enable acceptable reliability performance.

There are many radiation-induced soft and hard failure mechanisms that effect different types of circuit components in different radiation environments (terrestrial, avionics, space, nuclear reactors, nuclear blasts, etc.). It is beyond the scope of this chapter to deal with all of these, so our focus is limited to the primary SEE that induce soft errors in digital complementary metal-oxide-silicon (CMOS) circuits in the terrestrial environment. A single-event-upset (SEU) is the most common SEE and is usually a single-bit-upset. If the radiation event is of high energy, more than a single bit may be affected, creating a multi-bit-upset (MBU). While MBUs are usually a small fraction of the total observed SEU rate, their occurrence has implications for memory architectures utilizing error correction [3,4]. Another type of soft error can occur when the bit that is flipped is storing a critical system state, thus the error produces a large scale systemic malfunction involving failures in a large number of bits [5]. For example, a soft error in a register responsible for a chip reset or self-test routine that would be invoked when the error occurred in that register would lead to a chip failure since it would be reset or no longer be operating in the correct mode. This type of soft error, called a single-event-functional-interrupt (SEFI), directly impacts product reliability since each SEFI leads to a direct product malfunction as opposed to typical memory soft errors that may or may not effect the final product operation depending on the algorithm, data sensitivity, etc., Radiation events occurring in combinational logic result in the generation of single event transients (SET) which, if propagated and latched into a memory element, will lead to a soft error [6]. The SEE can also cause disruption of electrical systems indirectly by turning on vertical and lateral parasitic bipolar transistors between CMOS well and substrate—thereby inducing a high-current latch-up condition [7,8]. The primary difference between single event latch-up (SEL) and electrical latch-up is that the current injection that turns-on the parasitic bipolar elements is provided by a radiation event instead of an electrical event (current injection or voltage transient). Single event latch-up can be debilitating since its occurrence will necessitate a full-chip power-down to remove the condition. In modern micro-electronics, SEL is often non-destructive since current is limited by external circuit resistances; however, occasionally latent or catastrophic damage to metallization and junctions can lead to intermittent or hard errors.

31.2 The Terrestrial Radiation Environment

31.2.1 Ions in Matter

In order to appreciate the way in which the different terrestrial mechanisms produce disruptions in microelectronics, it is helpful to review some basic attributes of ions in matter. When an ion travels through a material, it loses its kinetic energy predominantly through interactions with the electrons of that material (3.6 eV per electron-hole pair in Silicon) and thus leaves a trail of ionization in its wake (ions can also interact directly with material nuclei but this reaction probability is usually significantly

lower than the electronic interaction). The higher the energy of the ion, the farther it travels before being "stopped" by the material. The distance required to stop an ion (its range) is a function of the ion's energy and the properties of the material (primarily the density) in which it is traveling. As the ions are slowed, they have more time to interact with the local electronic charge and, thus, their effect peaks near the end of their range. Particles are stopped when their kinetic energy is zero. The stopping power or linear energy transfer (LET) refers to the energy loss of the particle as a function of the distance traveled in that material. The LET (usually reported in units of MeV cm^2/mg) is a function of the ion's mass and energy and a function of the material's density (electronic structure and physical structure are also variables but to a lesser degree). A higher LET implies more energy deposited within a given volume. For example, the secondary particles produced by reactions of cosmic neutrons with silicon nuclei generate larger charge disturbances over shorter ranges than the relatively light alpha particles (helium nuclei) emitted from uranium and thorium impurities. In Figure 31.1 we show the LET of alpha particles, Silicon ions (in general the reaction products of cosmic neutron reactions with ^{28}Si will be lighter so this is an upper-bound), and Fe ions [9]. The LET can also be converted to charge generated/unit length (usually in units of fC/μm) as shown on the right-hand axis.

Integrating the LET over the length through which the event traverses the device volume gives an approximation of the charge collected by the device—a higher LET event will deposit more charge into a device and, thus, will have a higher probability of upsetting the device. While alpha particles and neutron reaction products are common in the terrestrial environment, heavy ions like iron are limited to the space environment. The LET in terrestrial environments is usually limited to <14 MeV cm^2/mg (the shaded region in Figure 31.1) which is equivalent to a peak charge generation of 145 fC/μm. Given that charge storage in the nodes of today's microelectronic circuits is in the order of 1–30 fC (for the high density low-voltage products) it is clear that even a single radiation event can easily disrupt most circuits.

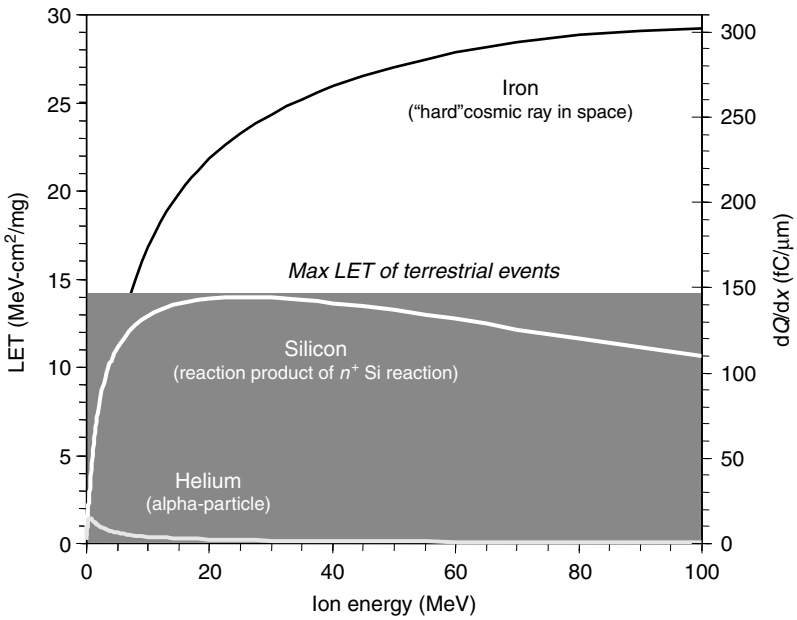


FIGURE 31.1 Linear energy transfer (LET) curves as a function of particle energy for several ions: alpha particles from trace uranium and thorium impurities and ^{28}Si ions as a worst-case example of secondary ions from the interaction of the cosmic background neutron flux and silicon nuclei. The LET of iron ions is shown for reference and is present in cosmic radiation encountered in space environments. Plots generated with SRIM 2006 (From Ziegler, J.F. and Biersack, J.P., The *Stopping and Range of Ions in Matter*, [version SRIM-2006.06 (c) 2006].)

31.2.2 High-Energy Cosmic Ray Neutrons

Primary cosmic rays are composed of galactic particles ($E \gg 1$ GeV) that are theorized to be left over from the Big Bang or remnants of Super Novae. There is an additional component from solar coronal mass ejections, flares, and prominences (usually $E < 1$ GeV). At the earth's outer atmosphere this deluge of energetic particles is composed of 92% protons, 6% alpha particles, and 2% gamma photons and heavier nuclei. Much of the energetic cosmic particle flux is trapped by the earth's magnetic field, but some of the particles make it into the atmosphere, where they react primarily via the Strong interaction and produce complex cascades of secondary particles. These, in turn, continue on deeper into the atmosphere, creating tertiary particle cascades, and so on. At terrestrial altitudes less than 1% of the primary flux reaches sea-level. The particle flux at sea-level is isotropic and composed of muons, protons, neutrons, and pions [10]. Altitude has the biggest effect on terrestrial neutron flux and consequently on the neutron-induced SEU rates. As altitude is increased, the decreased atmospheric density decreases the probability that the cosmic neutron will be absorbed before reaching the device. Over terrestrial altitudes from sea-level to 10 kft (~3000 m) the neutron flux increases nearly 20×. At avionics altitudes 28–60kft (10–20km) the neutron flux is a staggering 100–350× higher than sea-level flux. The earth's magnetic field strength has the second biggest effect on terrestrial neutrons. Since magnetic field lines enter the North and South poles nearly perpendicularly to the earth's surface, the magnetic shielding

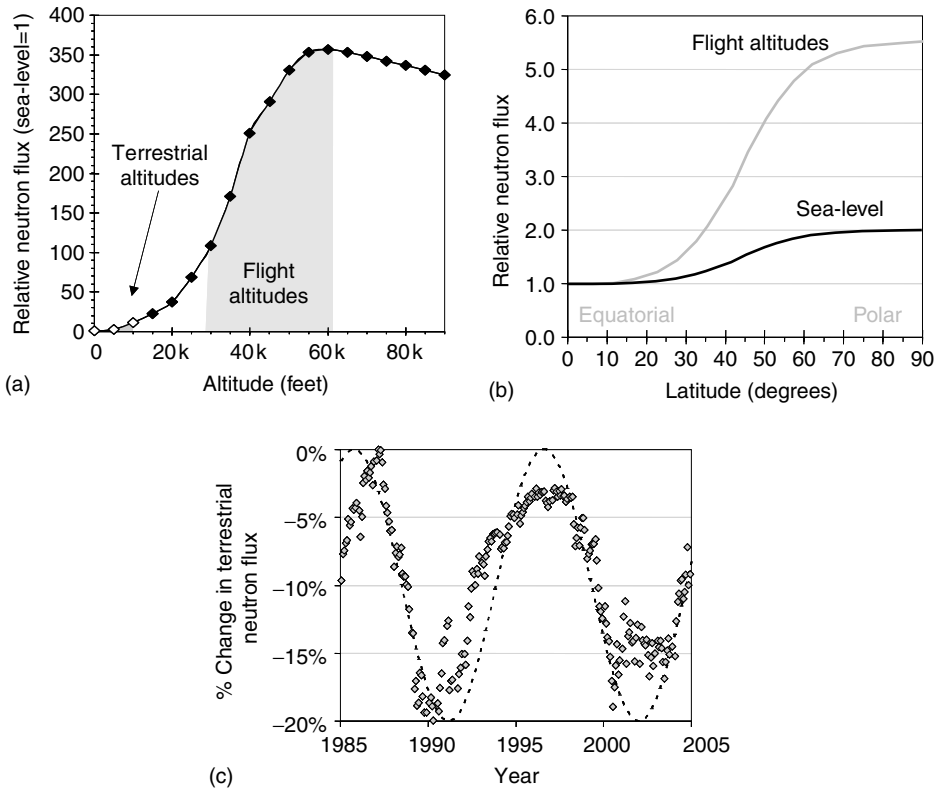


FIGURE 31.2 (a) Neutron flux variation as a function of altitude illustrates the 20× increase going from 0 to 14,000 ft. (From Normand, E., *IEEE Trans. Nucl. Sci.*, 43, 463, 1996.) (b) Neutron flux varies also with latitude, by about 2× from the equator to the poles. (From Normand, E., *IEEE Trans. Nucl. Sci.*, 43, 463, 1996.) (c) Neutron flux is also modulated by the 11-year solar activity cycle, by about 20%. (From Moscow Neutron Monitor web site: http://helios.izmiran.rssi.ru/cosray/main.htm [accessed on January, 2007].)

effect (geomagnetic rigidity) diverting or trapping cosmic protons is minimized at the poles. The neutron flux at high latitudes is thus significantly higher than at the equatorial regions where the magnetic field lines are parallel to the Earth's surface and more easily deflect or trap the protons. Terrestrial neutron flux varies with latitude and, to a lesser degree, longitude, by about 2×—with the neutron flux maximized at the poles. The final variable defining the neutron flux is solar activity. The solar activity cycle is an approximately 11-year variation in frequency and number of sunspots, coronal mass ejections, solar flares, and solar prominences. These solar cycle effects typically lead to about a 20% variation in neutron flux. The effect of altitude (20×), latitude (2×), and solar cycle (20%) on neutron flux are shown in Figure 31.2a, [11], Figure 31.2b [11], and 31.2c [12], respectively.

Although pions, muons, electrons, and protons of cosmic origin are capable of inducing charge disturbances, 90% of all events are caused by neutrons, and thus they are the dominant cosmic source of soft errors in the terrestrial environment. The terrestrial neutron spectrum spans from thermal energies to 1 GeV as shown in Figure 31.3 which shows the differential neutron flux at sea-level as a function of neutron energy [13,14]. Integrating from the lowest neutron energy capable of upsetting a device to the maximum neutron energy, the terrestrial neutron flux is found to be about 15 N/cm^2 h (at sea-level). Considering that the probability of interaction (reaction cross-section) is relatively low, it would seem that there are not enough neutrons to cause a significant problem, yet even a flux 15 N/cm^2 h is capable of inducing soft errors at a rate of many thousands of FIT in modern technologies.

Neutrons can interact both elastically and inelastically with device nuclei (predominantly ^{28}Si or ^{16}O in silicon-based semiconductor devices). In elastic reactions, in a similar fashion to billiard balls, an energetic neutron transfers some of its kinetic energy to the ^{28}Si or ^{16}O nucleus allowing the nucleus to be ejected from its lattice position within the material. The ion produces a high density of electron-hole pairs in the wake of its trajectory until it is eventually stopped. In inelastic reactions, the neutron actually gets absorbed into the nucleus leading to an instability that causes energetic secondary ions to be ejected from the nucleus [15,16]. Like the elastic recoil, these ions generate a high density of charge. Table 31.1 shows some of the lower energy reaction pathways for the n^{+28}Si reactions along with the required neutron energy to initiate the reaction. Since the secondary ions resulting from these neutron reactions such as Mg, Al, and Ne typically have high LETs

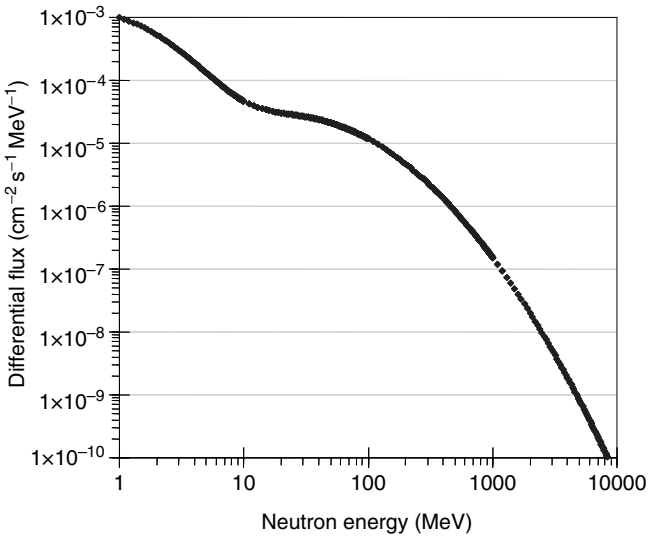


FIGURE 31.3 Differential neutron flux of terrestrial cosmic rays as a function of neutron energy at sea-level. (Adapted from Goldhagen, P., *Mater. Res. Soc. Bull.*, 28, 131–135, 2003.)

TABLE 31.1 Secondary Ion Reaction Products and Respective Reaction Threshold Energies for Some Primary $n^{+28}Si$ Inelastic Reactions

Reaction Product	Threshold Energy (MeV)
$^{25}Mg+\alpha$	2.75
$^{28}Al+p$	4.00
$^{27}Al+d$	9.70
$^{24}Mg+n+\alpha$	10.34
$^{27}Al+n+p$	12.00
$^{26}Mg+^{3}He$	12.58
$^{21}Ne+2\alpha$	12.99

(as shown in Figure 31.4), they generate hundreds of fCs of charge for every micron traveled, and with many CMOS devices operating on charge budgets of 20 fC or less, neutron reactions occurring in close proximity to device junctions will almost always induce soft errors. Additionally, certain SEE phenomena like MBU and SEL are rarely induced by alpha particles because the LET threshold for these types of events tends to be higher (in other words, these events typically require much larger charge disturbances to occur) than the peak LET for alpha particles. Thus MBU and SEL are typically induced by high-energy neutron reactions.

31.2.3 Alpha Particles from the Radioactive Decay of Impurities

In the late 1970s, alpha particles emitted by trace uranium and thorium impurities in packaging materials were shown to be the dominant cause of soft errors in dynamic random-access memory (DRAM) devices [17]. The alpha particle is composed of two neutrons and two protons, a doubly-ionized helium atom emitted from the nuclear decay of unstable isotopes. The most common source of alpha particles are from the naturally-occurring $^{238/235}U$ and ^{232}Th. A population of $^{238/235}U$ atoms in equilibrium emits

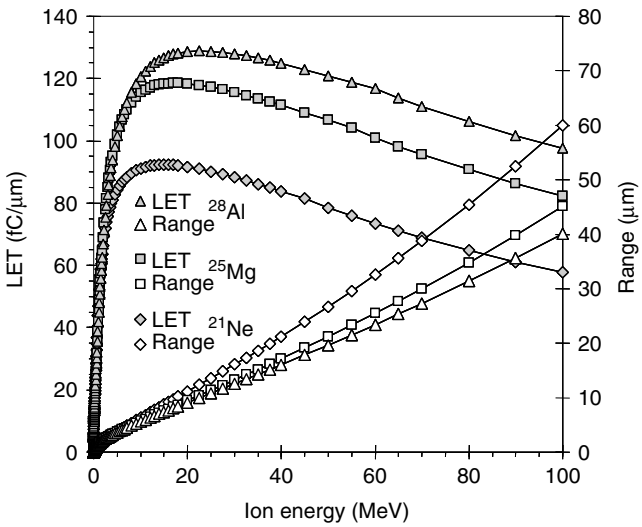


FIGURE 31.4 LET and range curves for several secondary ions emitted after inelastic $n^{+28}Si$ reactions as a function of ion energy in silicon. Note that all these ions are capable of generating high charge densities with peak LETs in the range of 80–130 fC/μm! Generated with SRIM 2006. (From Ziegler, J.F. and Biersack, J.P., The *Stopping and Range of Ions in Matter*, [version SRIM-2006.06 (c) 2006].)

sixteen different alpha particles at well-defined energies ranging from about 4 to 8 MeV, and an equilibrium population of ^{232}Th will emit six alpha particles from 4 to 9 MeV. Although the alpha emission occurs at discrete emission energies, the actual alpha particle fluxes incident on the device silicon in packaged components are defined by a broad energy spectrum spanning from 0 to 9 MeV. The broadening of the discrete emission energies occurs because the alpha-emitting impurities are generally uniformly distributed and travel different distances and lose different amounts of energy depending on their trajectories before being emitted from the material's surface. This situation is illustrated in Figure 31.5, showing both the discrete alpha emission spectrum and the simulated alpha spectrum emitted from a thick silicon-dioxide layer with a uniform distribution of ^{228}Th throughout its volume.

While alpha particles are directly ionizing, their LET is significantly lower than that of the secondary ions produced by high-energy neutron reactions. Alpha particles have a peak LET of about 16 fC/μm. Even the highest energy alpha particle emitted from U/Th impurities ($\sim$9 MeV) has a range in silicon of $<$100 μm as shown in Figure 31.6. Thus alpha particles from outside the packaged device are not a concern—only alpha particles emitted by the device materials and packaging materials need be considered. Since virtually all semiconductor materials are highly purified, the alpha-emitting impurities will generally not be in equilibrium. Alpha counting (where large area samples are counted for hundreds of hours by sensitive ionization detectors) must be used to determine the alpha emission since the exact nature of parent/daughter distributions is seldom known. To use mathematical parlance, a low concentration of uranium and thorium impurities is *a necessary requirement for low alpha emission but is not sufficient to guarantee low emission,* since in non-equilibrium situations higher activity daughters may be present that greatly increase the alpha emission rate. This situation was highlighted during investigations into eutectic lead solders (for flip-chip bumps) in which all radioactive impurities had been eliminated except the radioactive ^{210}Pb that was chemically inseparable from the ^{206,208}Pb. Since ^{210}Pb does not emit an alpha particle when it decays, initial alpha counting measurements revealed the solder to be emitting alpha particles at extremely low levels [18]. With the relatively short half-life of ^{210}Pb, a re-growth of the alpha emitter ^{212}Po (from the decay of ^{210}Pb$\Rightarrow$ ^{210}Bi$\Rightarrow$ ^{210}Po) occurred, and within a few months the solder alpha emission was 10$\times$ higher than initial measurements indicated.

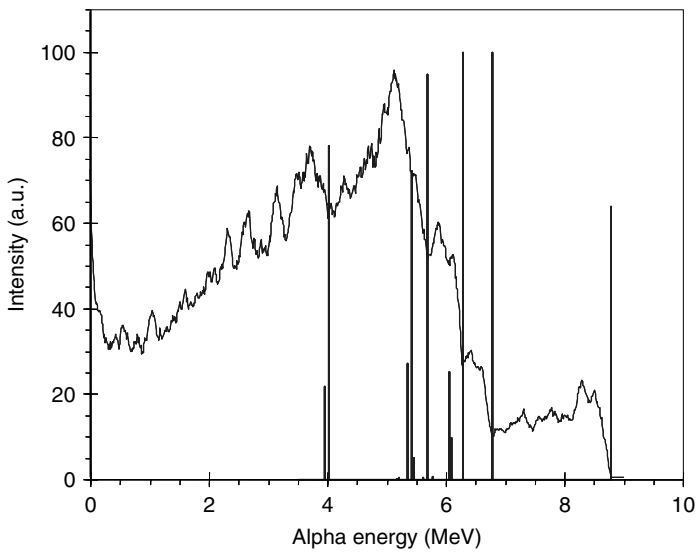


FIGURE 31.5 Simulated discrete energy alpha energy spectrum obtained from a surface concentration of ^{232}Th and simulation of spectrum emitted from a thick sample of SiO_2 assuming a uniform distribution of ^{232}Th impurities within the material volume. Note the broadening of discrete emission lines due to energy loss in the SiO_2.

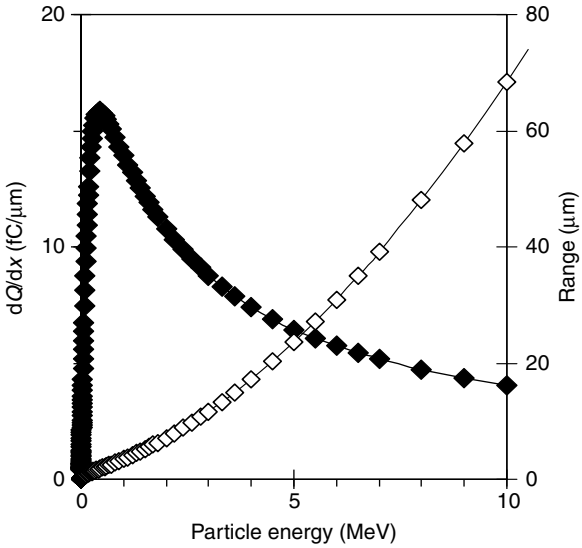


FIGURE 31.6 LET (solid diamonds) and range (open diamonds) of an alpha particle in silicon as a function of its energy. Note that all alpha particles are stopped within 100 μm and thus only alpha particle sources within the packaged device need be considered. Plot generated with SRIM 2006. (From Ziegler, J.F. and Biersack, J.P., The *Stopping and Range of Ions in Matter*, [version SRIM-2006.06 (c) 2006].)

Impurity analysis would not have revealed any issue either, since problematic alpha emission levels can occur even when the ^{210}Pb level is far below the detection limits of any spectroscopic method. A ^{210}Pb/^{208}Pb impurity ratio as low as 10^{-17} is still high enough to cause alpha SER problems. For advanced electronics most manufacturers have moved to the use of low or ultra-low alpha (ULA) materials. For a material to be classified as ULA (ULA implies and emission at or below 0.002 α/h cm^2) the $^{238/235}$U and ^{232}Th impurity content must be below about one part-per-ten billion!

In CMOS devices where the semiconductor manufacturing and packaging materials have been purified and screened such that together they contribute <0.002 α/h cm^2, the fraction of soft errors from alpha particles will be $<50\%$ in most cases (the actual alpha SER will depend on the voltage of operation, number of metal layers, and the emission of all materials combined). At this point further emission reduction becomes prohibitively expensive while providing diminishing returns since the SER is dominated by cosmic background radiation. As shown in Table 31.2, the chip-level alpha flux is usually dominated by emission from the packaging materials, as the purity of semiconductor materials and the semiconductor manufacturing process is usually significantly better than that of the packaging materials and processes.

TABLE 31.2 Measured Alpha Particle Emission Levels from Various Chip and Packaging Materials Used in Semiconductor Manufacturing

Chip/Package Materials	Alpha Emission (α/cm^2 h)
Bare silicon wafer	<0.00004
Fully processed 8-level Cu wafer	<0.00028
Package underfill compound	<0.001
Package silica-based mold compound	<0.001
Package lead-based solder bumps	7 to <0.002

31.2.4 Low-Energy Cosmic Ray Neutrons and ^{10}B

The third significant source of ionizing particles in electronic devices is the secondary radiation induced from the interaction of low-energy cosmic ray neutrons and boron. First proposed to be a potential issue in spacecraft due to the boron in packaging materials [19] and mentioned as a potential risk when boron was used as a substrate dopant [20], boron was actually demonstrated to be a dominant source of soft errors in devices using boron-doped phosphosilicate glass or BPSG [21,22] insulation layers in integrated circuits. While the previous discussion on cosmic neutrons focused on the high-energy reactions, here the focus is on very low-energy neutrons ($\ll$1 MeV).

Boron is composed of two isotopes, ^{11}B (80.1% abundance) and ^{10}B (19.9% abundance). The ^{10}B is unstable when exposed to neutrons and breaks into ionizing fragments shortly after absorbing a neutron (the ^{11}B also reacts with neutrons; however, its reaction cross-section is nearly a million times smaller, and its reaction products, gamma rays, are generally incapable of producing soft errors). At 3838 barns (1 barn$=10^{-24}$ cm^2 per nucleus), the thermal neutron capture cross-section of ^{10}B is 3–7 orders of magnitude higher than most other isotopes present in semiconductor materials. Unlike most isotopes which emit innocuous gamma photons after absorbing a neutron, the ^{10}B nucleus breaks apart with an accompanying release of energy in the form of a ^{7}Li recoil nucleus and an alpha particle (a prompt gamma photon is also emitted from the lithium recoil soon after fission occurs). The alpha particle and the lithium nucleus are emitted in opposite directions to conserve momentum—thus in a surface layer above the silicon, it is highly likely that one of the two ions will have a trajectory towards the active devices. The lithium nucleus is emitted with a kinetic energy of 0.840 MeV 94% of the time and 1.014 MeV 6% of the time. The alpha particle is emitted with an energy of 1.47 MeV. As the lithium recoil and alpha particle emitted have a peak LET of 25 and 16 fC/μm, respectively, both events are capable of inducing soft errors in electronic devices. The LET and range of lithium recoils and alpha particles is shown in Figure 31.7. Since the ^{10}B neutron

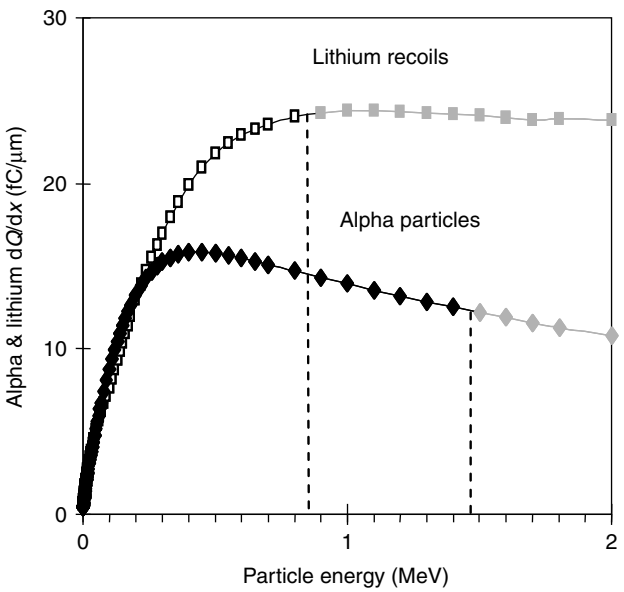


FIGURE 31.7 LET for the lithium recoil (solid diamonds) and alpha particle (open squares) reaction products of the n+^{10}B reaction as a function of their energy. Note that both these ions have relatively low energy and thus only reactions that occur within a few microns of the active device areas need be considered. Plot generated with SRIM 2006. (From Ziegler, J.F. and Biersack, J.P., The *Stopping and Range of Ions in Matter*, [version SRIM-2006.06 (c) 2006].)

capture cross-section decreases rapidly as neutron energy is increased, low-energy neutrons (thermal/epi-thermal) dominate the reaction. However, since the range of the alpha particle and lithium recoil is $<3\,\mu m$, only ^{10}B in close proximity to the silicon substrate is a threat. ^{10}B in the packaging materials is no threat at all because the secondary ions do not have enough energy to reach the chip surface. For conventional BPSG-based semiconductor processes, the BPSG is the dominant source of boron reactions and in some cases can be the primary cause of soft errors.

Soft errors in semiconductor devices in the terrestrial environment are due to three mechanisms; terrestrial cosmic radiation in the form of high-energy neutrons causing nuclear reactions with device materials whose secondary products are highly ionizing, directly ionizing alpha particles emitted from the radioactive impurities in the device materials, and the nuclear reaction of low-energy neutrons from the same cosmic background with ^{10}B in devices, producing ionizing alpha particles and lithium recoils. In advanced devices where boron-doped glass is not used and all chip materials are screened to ensure that their emission is ULA ($<0.002\ \alpha/cm^2\ h$), the high-energy cosmic ray neutrons are the dominant cause of soft errors with alpha particles typically contributing $<50\%$ of the total soft errors.

31.3 Radiation Effects on Semiconductor Devices

The most sensitive semiconductor device structure is the reverse-biased junction. In the worst case, the junction is floating (as in DRAMs, dynamic logic circuits, and some analog designs) and is extremely sensitive to any charge collected from a radiation event reduces the stored signal charge. Furthermore, the n^+/p node is worse than its compliment because the collection of electrons (which will compensate the inversion hole charge) is more efficient due to the greater mobility of electrons as compared with holes. In addition, since most manufacturers use p-type substrates, collection can occur deep within the bulk (as opposed to p^+/n devices that are formed in n-wells in the p-substrate where a portion of the charge transient will be removed by the well contact).

Consider the effect of an ion on a reverse biased n^+/p junction with a positive voltage on the n^+ node. At the onset of the radiation event (Figure 31.8a), a cylindrical track of electron-hole pairs with a sub-micron radius and very high carrier concentration is formed in the wake of the energetic ion's passage ($<$ 0.1 ps). When the resultant ionization track traverses or comes close to the depletion region, carriers are then rapidly collected by the electric field creating a large current/voltage transient at that node (Figure 31.8b). A notable feature of the event is the concurrent distortion of the potential into a funnel shape [23]. This funnel greatly enhances the efficiency of the drift collection by extending the high-field depletion region deeper into the substrate. The size of the funnel is a function of substrate doping—the funnel distortion increasing for decreased substrate doping. This "prompt" collection phase is completed within a nanosecond and followed by a phase where diffusion begins to dominate the collection process (Figure 31.8c). Additional charge is collected as electrons diffuse into the depletion region on a longer time scale (hundreds of nanoseconds) until all excess carriers have been collected, recombined, or diffused away from the junction area. The diffusion process is much slower and typically the total charge collected from diffusion is usually significantly less than that collected initially by prompt collection in the case of advanced technologies. Events that occur some distance from the sensitive node, though, can be dominated by diffusion but ultimately end with much less total collected charge. The resultant current transient on the node is shown in Figure 31.8d (It should be noted that the collected charge induces both a current and voltage transient at the node). In products there is never a single isolated node but a "sea of nodes" in close proximity. Thus it is important to note that charge-sharing among nodes will occur and that parasitic mechanisms such as unintentional bipolar action and current shunts can also be major features of collection in real devices.

The magnitude of the collected charge (Q_{coll}) depends on a complex combination of factors including the size of the device, biasing of the various circuit nodes, substrate structure, device doping, the type of ion, its energy, its trajectory, the initial position of the event within the device, and the state of the device.

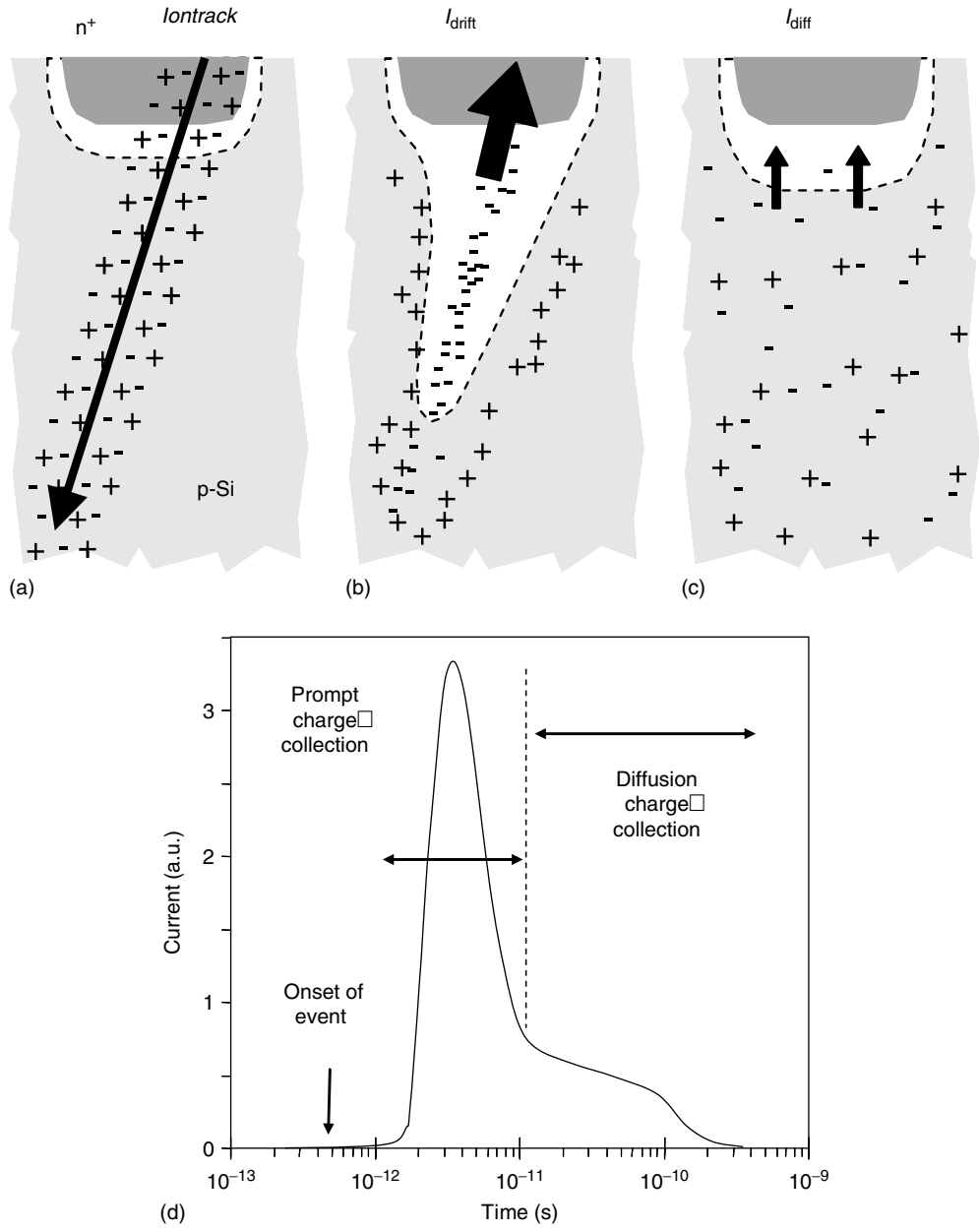


FIGURE 31.8 Diagram of the charge generation and collection process occurring after an ion strikes a junction. (a) A dense non-equilibrium cylindrical distribution of electron-hole pairs is left in the ion's wake. (b) As the distribution segregates, a potential distortion, or "funnel" is formed, accelerating the collection of electrons at the junction by drift. (c) Eventually the funnel collapses and collection continues at a slower rate by diffusion. (d) The current transient produced by such an event will have a sharp peak during the drift collection phase, followed by a wider region dominated by diffusion collection.

However Q_{coll} is only half the story, as the device's sensitivity to this excess charge needs to be taken into account as well. This sensitivity is defined primarily by the node capacitance, operating voltage, and the strength of feedback transistors, all defining the amount of charge, or critical charge (Q_{crit}) required to trigger a change in the data state. The response of the device to the charge injection is dynamic and dependent on the magnitude and the temporal characteristics of pulse, and thus Q_{crit} is not constant but depends on the radiation pulse characteristics and the dynamic response of the circuit itself, making the effect difficult to model [24]. For simple isolated junctions (like DRAM cells in storage mode) a soft error will be induced when a radiation event occurs close enough to a sensitive node such that $Q_{coll} > Q_{crit}$ and Q_{crit} have the simple form:

$$Q_{crit} = C_{node} V_{node} \tag{31.1}$$

Conversely, if the event results in a $Q_{coll} < Q_{crit}$ then the circuit will survive the event and no soft error will occur. It should be noted that Equation 31.1 holds for DRAM at low frequencies where cell hits dominate the SER. If a cell collects more charge than the critical charge for the circuit, the "1" level will be destroyed. Cell hits will virtually always be the changing of a "1" data state into a "0" since for the zero case the junction is in a relaxation state and cannot be easily be disrupted. As the frequency is increased, the circuit will be refreshed more often, and bitline events will become more likely. It should be noted that for bitline hits, collection in the bitline diffusion and sense amplifier junctions must also be considered. The bitline and sense amplifier are only sensitive during the sensing portion of the timing when the bitlines are floating (not driven) and the sense amplifier is sensing. Since the sense timing is generally defined by parasitic capacitances/resistances, increasing the frequency increases the percentage of time the sense amplifier is sensitive to a radiation event. Thus SER due to bitline/sense amplifier hits increases linearly with increasing frequency while cell SER stays constant or decreases slightly. For modern high-speed DRAMs a large portion of the SEU is due to bitline or sense-amp hits. Any charge deposited on the bit-line during sensing may erroneously alter the output of the sense amplifier at the end of the sense interval. Depending on which bitline is hit, the error maybe a "1" or "0".

In static random access memory (SRAM) or other logic circuits where there is active feedback, an additional term is necessary to comprehend the speed with which the circuit can react—slower speeds allow more time for the feedback circuit to restore the corrupted node value and thereby reducing the probability of a soft error. This additional term tends to increase the effective Q_{crit} and can be expressed as:

$$Q_{crit} = C_{node} V_{node} + \tau_{switch} I_{restore} \tag{31.2}$$

The standard 6T SRAM cell is composed of two transistors to allow access to the storage cell via the word-lines. The storage cell is composed of two P-channel metal-oxide semiconductor (PMOS) and two N-channel metal-oxide semiconductor (NMOS) transistors forming two cross-coupled inverters as illustrated in Figure 31.9. Note that the SRAM cell is in storage mode. In other words, the word lines are low so that the cell is isolated from the bit-lines—during a read or write operation, the word-lines would be brought high so that the data state of the cell could be written from or read to the bit-lines. The regenerative feedback loop thus maintains the data state of the cell as long as power is applied. If an ion traverses the node storing the "1" data state the resulting e–h pairs produced will be separated by the local fields and excess electrons will be collected causing a rapid drop in the stored voltage of the left node storing the "1" data state. As the node voltage drops, hole current from the left-hand PMOS will start to compensate. Whether or not this event causes the SRAM bit to flip depends on, whether or not the PMOS can supply enough current to compensate the current induced by the event before the cell itself flips to the opposite data state. Note that as the node voltage drops on the left side, this has consequences for the right side as the right-hand PMOS starts to turn-on while the right-hand NMOS is turning off. This further aggravates the situation since this will tend to turn-off the left PMOS while turning on the left NMOS, actually helping to bring the left node down and the right-hand node high. If the hole current from the left PMOS cannot quench the excess charge before the left storage node climbs below some

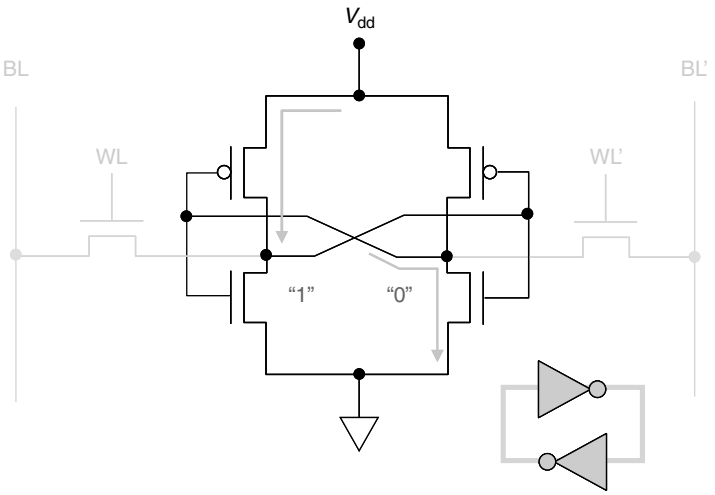


FIGURE 31.9 A standard 6T static random access memory (SRAM) cell in storage mode (word lines are low thereby isolating the cell from the bitlines). Note that the cell is actually composed of two cross-coupled inverters so that the output of one inverter is driving the input of the other inverter in a regenerative feedback loop.

critical low voltage value, switching will occur and an error will be produced. The soft error process in Flip-flops (which are usually created from a master and slave stage, each similar to a single SRAM cell) and latches is similar to the SRAM process described above.

31.4 Technology Scaling Trends

31.4.1 Memory SER Sensitivity

To create the functionality provided by today's electronic systems and appliances several distinct components must be integrated together. At the core of each system is a microprocessor or digital signal processor with large embedded memories (usually SRAM) interconnected with a slew of peripheral logic. In larger systems discrete main memory (usually DRAM) is also used. Finally, all systems have some analog or digital input/output components to allow the device to respond and interact with the outside world. The SER of these various chip components behaves differently as the respective technologies are scaled.

It is somewhat ironic that soft errors were first discovered to be a problem in DRAM, because after many generations, it is currently one of the more robust electronic devices. Dynamic random-access memory bit SER was high when manufacturers used planar capacitor cells that stored the signal charge in 2-D, large-area junctions because these were very efficient at collecting radiation-induced charge. To address pause-refresh and soft error problems while increasing packing density, DRAM manufacturers migrated to 3-D capacitor designs that significantly increased the Q_{crit} while greatly reducing junction collection efficiency by eliminating the large storage junction in silicon [25]. There are two predominant DRAM cell designs. One is the trench cell in which the 3-D capacitor structure is formed by milling a deep cylindrical trench in the substrate. The trench is then lined with a thin high-dielectric constant material followed by a poly-silicon deposition to fill the trench. Increased capacitance for each node is achieved by the use of high-k materials, thinner dielectrics, and deeper trenches. The other archetype of modern DRAM cells is the stack/fin capacitor which is formed above the transistors and has a "coffee cup" like shape with inter-digitated layers to further boost the capacitance. To further increase capacitance while scaling to smaller dimensions, the height of the stack/fin is increased, the polysilicon electrodes are roughened with specific etches, and extra fins/plates are added. A junction is used in both

trench and stack-fin designs as one side of the pass transistor. Since this junction is of minimum size, its collection efficiency is far below that of the old planar type DRAM. Additionally, collection efficiency decreases with the decreasing volume of the junction (junction/well doping also play a role) while the cell capacitance remains relatively constant with scaling since it is dominated by the external 3-D capacitor cell.

These DRAM device scaling trends are illustrated in Figure 31.10a, along with DRAM cell voltage scaling. The modest voltage scaling has reduced Q_{crit} but with the concurrent aggressive junction volume scaling, a much more significant reduction in collected charge is observed. The net result to DRAM SER performance is shown in Figure 31.10b, with the SER of a DRAM single bit shrinking about 4–5× per generation. While DRAM bit SER has been reduced by more than 1000 times over seven generations, the DRAM system SER has remained essentially unchanged. System requirements have increased the memory density (bits/system) almost as fast as the SER reduction provided by technology scaling. Thus, DRAM system reliability has remained roughly constant over many generations. So contrary to the popular misconception that DRAM SER is problematic, undoubtedly left over from the days when DRAM designs utilized planar cells, DRAM is one of the more robust devices in terms of soft error immunity.

In contrast early SRAM was more robust against SER because of high operating voltages and the fact that data in an SRAM is stored as an active state of a bi-stable circuit made up of two cross-coupled inverters, each strongly driving the other to keep the SRAM bit in its programmed state. The Q_{crit} for the SRAM cell is largely defined by the charge on the node capacitance as with DRAM but with a dynamic second term related to the drive capability of the transistor keeping the node voltage at the proper value—the stronger the transistor, the more charge that must be collected for the node voltage to reach the switching threshold. With technology scaling, the SRAM junction area has been deliberately minimized to reduce capacitance, leakage, and cell area, while, simultaneously, the SRAM operating voltage has been aggressively scaled down to minimize power. These device scaling trends are shown in Figure 31.11a. With each successive SRAM generation, reductions in cell collection efficiency due to shrinking cell depletion volume have been cancelled out by big reductions in operating voltage and reductions in node capacitance. It can be seen that SRAM single bit SER initially was increasing with each successive generation, particularly in products using BPSG as illustrated in Figure 31.11b. Most recently as feature sizes have been reduced into the deep submicron regime (<0.25 μm), the SRAM bit SER has

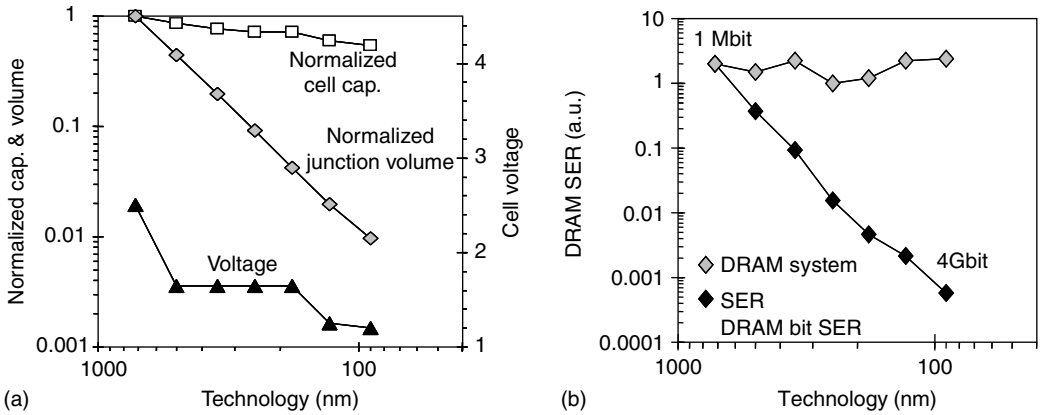


FIGURE 31.10 (a) Key dynamic random access memory (DRAM) technology scaling parameters; normalized cell capacitance, normalized junction volume, and cell voltage as a function of technology node (generation), and (b) DRAM single bit soft error rate (SER) and system SER as a function of technology node (generation). Single bit SER allows a direct sensitivity comparison across technologies while the system SER accounts for increased memory density (number of bits) as technologies are scaled.

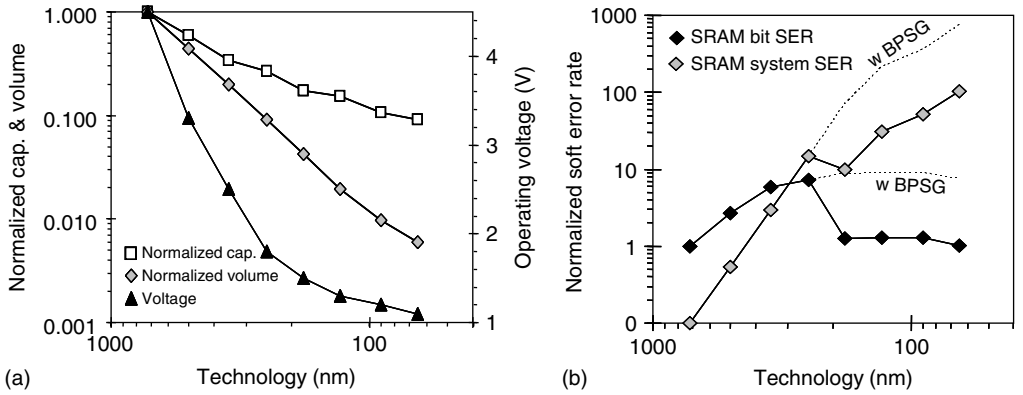


FIGURE 31.11 (a) Some key SRAM parameters, normalized storage node capacitance, normalized junction volume, and voltage as a function of technology node (generation), and (b) SRAM single bit and system SER as a function of technology node. Single bit SER allows a direct sensitivity comparison across technologies while the system SER accounts for increased memory density (number of bits) as technologies are scaled. Note the reduction in SER following the 0.25 μm node due to borophosphate silicate glass (BPSG) elimination (dotted lines show SER with BPSG present).

saturated and may even be decreasing. This saturation is primarily due to the saturation in voltage scaling, reductions in junction collection efficiency, and increased charge sharing due to short-channel effects with neighboring nodes. Scaling also implies increased memory density, so the saturation in SRAM bit SER does not translate into saturation in the SRAM system SER. The exponential growth in the amount of SRAM in microprocessors and digital signal processors has led the SER to increase with each generation with no end in sight. This trend is of great concern to chip manufacturers since SRAM constitutes a large part of all advanced integrated circuits today.

31.4.2 Sequential/Combinational Logic SER Sensitivity

A computer's discrete and embedded SRAM and DRAM memories would be useless without the peripheral logic that interconnects them. While less sensitive than SRAM, logic devices can also experience soft errors [26–29]. Sequential logic elements include latches and flip-flops used to hold system event signals and to buffer data before it goes in or out of the microprocessor and to interface to combinational elements that perform logical operations based on multiple inputs. The SER of these devices and its impact on the system are much harder to quantify since their period of vulnerability (when they are actually doing something critical in the system versus simply waiting) varies widely depending on the circuit design, frequency of operation, and the actual algorithm being executed. Flip-flops and latches are fundamentally similar to the SRAM cell in that they use cross-coupled inverters to store the data state. However, they tend to be more robust because they are usually designed with at least two stages and often with larger transistors, which can more easily compensate for spurious charge collected during radiation events. Ultimately, the reliability concern with sequential and combinational logic circuits is that, like SRAM, their SER sensitivity is also increasing with scaling as illustrated in Figure 31.12. The black diamonds represent the native SRAM bit SER. The shaded region at the bottom of the plot represents the effective SRAM bit SER with error correction. It should be noted that error correction can offer even more significant improvement, but we assume a long data residency and a minimum multiplex factor (physical distance between bits in the same correction word). With frequent read-write accesses and/or scrubbing, the actual corrected SER can be significantly lower than indicated here. The plot with the triangles shows simulated and experimental data representing the sensitivity of sequential logic on a per-bit or per-flip-flop basis. Clearly as silicon technology is scaled, the SER in

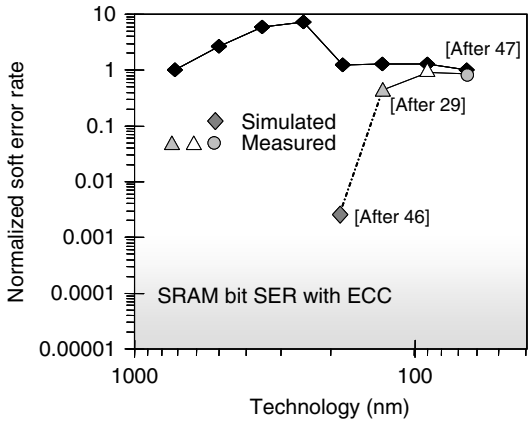


FIGURE 31.12 Comparison of SRAM single bit SER with flip-flop/latch SER obtained from test structures, product characterizations, and simulations. The gray region at the bottom of the plot represents the effective bit failure rate of SRAM with error correction employed.

sequential logic becomes non-negligible. Soft errors in logic are particularly a concern in high-reliability systems whose memory has been protected by error correction where the peripheral logic failure rate may be the dominant reliability failure mechanism.

In a combinational circuit where the output is based on a logical relation to the inputs (with no capability for retention), if enough radiation-induced charge is collected, a short-lived transient in the output will be generated (SET) [30]. If this radiation-induced "glitch" is actually propagated to the input of a latch or flip-flop during a latching clock signal, the erroneous input will be "latched" and will be stored. For older technologies the SET could not propagate since it usually could not produce a full output swing and/or was quickly attenuated due to large load capacitances and large propagation delays. In advanced technologies where the propagation delay is reduced and the clock frequency is high, the SET can more easily traverse many logic gates, and the probability that it is latched increases. Single event transients-induced soft errors are not expected to become an issue until the 65 nm technology node or beyond. It should be noted that once an SET can propagate freely, synchronous, and especially asynchronous (self-clocked) circuits would be extremely sensitive to such events. In technology nodes beyond 90 nm and at high product operating frequencies, there is increased risk that a large fraction of observed soft failures will be related to latched SET events [31].

31.5 Commercial Mitigation Techniques

31.5.1 Source Mitigation Techniques

Having made the decision that a particular product's SER is too high, mitigation strategies need to be considered. Unlike military or space electronics that may place a high premium on radiation robustness, commercial mass-market products can seldom employ expensive custom solutions and must always balance low cost, high performance, early time-to-market, and reliability. The first and most obvious way to mitigate a soft error problem is to get rid of the radiation sources that cause them.

To reduce alpha particle emissions, semiconductor manufacturers use extremely high purity materials and processes, production screening all materials with low background alpha emission measurements. By controlling impurity levels to fractions of parts-per-billion (this is necessary but not sufficient since the materials are typically not in equilibrium, so high-intensity daughter products far below detection limits can still cause alpha particle issues) coupled with direct alpha counting to

confirm that alpha emission is $<0.002\ \alpha/cm^2$-h (ULA), the fraction of SER from alphas can be reduced to $<50\%$ of the total SER.

Another method of reducing alpha particles or to further reduce the alpha SER, one can layout the chips and design the package such that the materials with the highest alpha emission are kept physically separated from sensitive circuit components. In the case of flip-chip with solder balls in close proximity to active silicon, bump keep-out zones are often used so that no solder bumps are allowed within a certain distance of sensitive devices and circuits. For advanced technologies, the embedded SRAM is often much more sensitive than the surrounding logic and thus typically keep-out zones are used in SRAM areas. This is an expensive procedure in terms of design routing as it can increase die size by up to 15%–20% depending on the size of sensitive areas and the number of bumps.

One last solution frequently employed to shield the high alpha emission from packaging materials is to coat the chip with a thick polyimide layer prior to packaging. In general, if a polyimide shield is to be used, it should be thicker than the maximum range of the most energetic alpha particle expected. This results in a required polyimide thickness of $\sim$35–50 μm (depending on the number of metal layers) that is not really practical for most situations, so accurate characterization or 3D simulation should be used to determine the thickness which reduces the flux to the desired level while still maintaining manufacturability. In general a polyimide thickness beyond 10 μm is hard to manufacture since multiple deposition and curing steps are required.

The SER due to the activation of ^{10}B in BPSG can be mitigated in several ways [32]. The first and most direct way is to simply eliminate BPSG from the process flow. Due to the limited range of the alpha and lithium recoil emitted, typically only the first and second levels of BPSG need be replaced with a dielectric free of ^{10}B. In cases where the unique reflow and gettering properties of boron are needed, the regular BPSG process can be replaced by a process that uses isotopically enriched $^{11}BPSG$ without changing the physical or chemical properties of the film and without the requirement for new equipment or processing steps. If the process cannot be changed at all, using a package with low energy neutron shielding is a straight forward approach (mold compound with BPSG filler or some other neutron absorbing material instead of normal silica will mitigate the effect) [33].

While large reductions in SER are possible either by removing the sources of or shielding the ^{10}B reaction products and alpha particles, a large portion of the high-energy cosmic neutron flux will reach the devices and cause soft errors. So ultimately, if ^{10}B has been removed from the process and alpha sources minimized to ULA levels, the product SER will be limited by the high-energy cosmic background neutron radiation. If the soft failure rate is still unacceptable at this point, process technology or design intervention is required.

31.5.2 Process Technology Mitigation Techniques

The remaining SER can be addressed, to some extent, by process and technology choices. Substrate structures or doping profiles that minimize the depth from which carriers can be collected can have an impact on reducing Q_{coll} thus reducing SER. CMOS with deep-tank or buried implants that locally increase the substrate doping can improve SER by reducing the size of the funnel formed and by increasing substrate charge collection, thereby reducing the amount collected by sensitive nodes. This improvement comes at the cost of an additional implant layer and a thermal cycle. CMOS with actively biased active wells provide reduced SER sensitivity by reducing the charge collected by reverse biased drain nodes, since much of the charge will be collected at well-contacts. This improvement does however come at a cost. From a fabrication point of view, an additional mask/implant layer is needed along with a thermal cycle. From the performance point of view, since input parasitic capacitance increases with the isolated well, speed is decreased, circuit area is increased to accommodate the well, and barrier lowering by reduced well potentials can lead to an increased risk of parasitic bipolar action. In DRAM, multiple-well isolation has been used to reduce charge collection as well as to improve cell leakage and charge retention characteristics. Well-based mitigation technologies have also been suggested for CMOS logic [34] including buried layers, guard-rings, and the use of epitaxial substrates.

These techniques typically yield less than 3× improvements in SER and can occasionally lead to increased sensitivity.

Substrates incorporating a very thin silicon layer on a thicker layer of buried oxide (silicon on insulator—SOI) also have been shown to reduce SER sensitivity [35–37] as compared with bulk silicon. Reduction of substrate collection efficiency is the key parameter that reduces SER in SOI devices. The best partially depleted SOI technologies seem to offer about 5–8× improvement in SER robustness as compared with bulk (130–90 nm nodes) as shown in Figure 31.13a and Figure 31.13b for alpha and neutron SER, respectively, [38]. The negative aspect of partially-depleted SOI is that to reduce parasitic bipolar effects, body ties are required, increasing circuit complexity and layout area. Since SER is dominated by the Bipolar Junction Transistor (BJT) turn-on in partially depleted silicon on insulator (SOI), the SER increases with temperature as the bipolar gain increases. Fully depleted SOI can provide much better SER immunity since floating body effects are mitigated and short-channel and dopant variation effects are also eliminated. Elimination of the floating body effect means that the parasitic BJT can no longer be formed in fully-depleted SOI, so the probability of SEU or SEL due to a radiation induced BJT turn-on is eliminated. Ultimately, improvements garnered by substrate engineering provide a limited path for mitigating soft errors considering that for high reliability applications, reductions of 1000–10,000× in the intrinsic SER are needed. For the majority of process technology solutions the SER is reduced by $<100\times$ at the expense of additional process complexity, yield loss, and substrate cost.

Due to how data is stored in the non-volatile devices, floating charge on an isolated gate in the case of flash memories and a polarization state in the ferroelectric material of an Ferroelectric Random Access Memory (FRAM) the SER sensitivity can be extremely low. In general non-volatile memories such as flash and FRAM are immune to SEU events under terrestrial conditions while in retention mode. The SER of these devices is generally defined by external CMOS sensitivity and how often the flash/FRAM memory is accessed. Single-event-upset and SEFI have been observed in non-volatile memories (10× to 1000× lower than typical SRAM) [39,40].

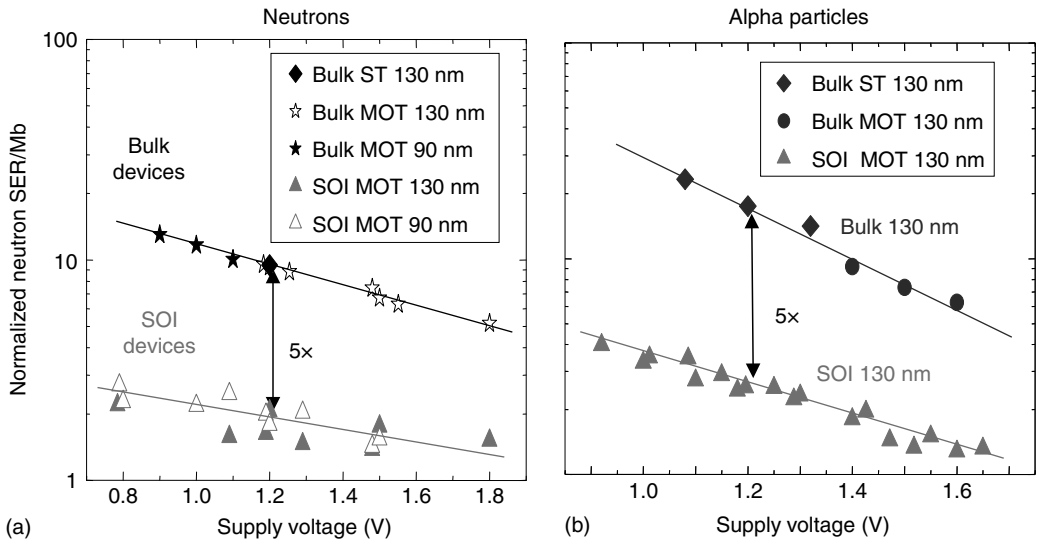


FIGURE 31.13 (a) Neutron and (b) alpha particle SER from bulk and partially-depleted silicon on insulator (SOI) SRAM for two technology nodes as a function of core voltage. In this implementation the use of SOI substrates yielded a 5× improvement in SER. (From Roche, P., Gasiot, G., Forbes, K., O'Sullivan, V., and Ferlet, V., *IEEE Trans. Nucl. Sci.*, 50, 2046–2054, 2003.)

31.5.3 Design Mitigation Techniques

Radiation sensitivity can be reduced significantly by design and layout changes. Any change which increases Q_{crit} while maintaining or reducing Q_{coll} will improve the SER performance of a device. For example, a typical high-density SRAM cell consists of six transistors; two allowing data to be read and written to-and-from the cell and four transistors making up the two cross-coupled inverters responsible for maintaining the data state. Q_{crit} is a function of the storage node capacitance and voltage and of an additional term for the restoring charge supplied by the pull-up/pull-down transistor. This restoring term is proportional to the switching time of the cell and the current provided by the load transistor. Increasing the current drive of the load transistors and/or increasing the switching time of the SRAM cell will increase the robustness of the cell against corruption. Thus Q_{crit} can be increased significantly if additional or larger drive transistors are added so that a larger restoring current can be provided during a radiation-induced transient. Resistance can also be added between the two inverters so that the time to flip the cell is increased [41], thus effectively allowing the pull-up/pull-down transistors more time to restore the data state (this approach affects the write-time of the cell and in high-speed technologies is not a realistic solution). Resistively hardened cells have been used extensively in the past to make SRAMs rad-hard. Basically the scheme is simple, with decoupling resistors used to slow the regenerative feedback response of the cell so that pull-up/pull-down transistors have time to restore the node voltage before a flip occurs. The SRAM response is thus slowed so that it cannot respond to the sub-nanosecond pulse induced by a radiation event. This approach is no longer reasonable for commercial applications with clock periods approaching or below 1 ns, since at this point adding enough resistance to filter out SEU will also constrain the operating frequency.

31.5.4 System-Level Redundancy Techniques

By far the most effective method of dealing with soft errors in memory components is by employing additional circuitry for error detection and/or correction. In its simplest form, error detection consists of adding a single bit to store the parity (odd or even) of each data word (regardless of word-length). Whenever data is retrieved, a check is run comparing the parity of the stored data to its parity bit. If a single error has occurred, the check will reveal that the parity of the data does not match the parity bit. Thus the parity system allows for the detection of a soft error for a minimal cost in terms of circuit complexity and memory width (only a single bit is added to each word) as illustrated in Figure 31.14a. The two disadvantages of this system is that the detected error cannot be corrected, and if a double error has occurred, then the check will not reveal that anything is wrong, since the parity will match. This is true for any even number of errors. For example, if the data was stored with odd parity, the first error changes the odd parity to even parity (detectable error), but the second error changes the parity back to odd (non-detectable error).

In order to address these short-comings, error detection and correction or error correction codes (EDAC or ECC) are employed. Typically error correction is achieved by adding extra bits to each data vector encoding the data so that the "information distance" between any two possible data vectors is, at least, three. Larger information distances can be achieved with more parity bits and additional circuitry—but in general the single error correction double error detection schemes are favored. In these systems, if a single error occurs (a change of plus or minus one in information space) there is no chance that the corrupted vector will be mistaken for its nearest neighbors (since the information distance is three). In fact, if two errors occur in the same "correction word" a valid error vector will be produced as illustrated in Figure 31.14b. The only limitation is that with two errors the error vector will not be unique to a single data value, thus only detection of double-bit errors is supported. For a 64-bit wide memory, eight correction bits are required to allow two errors to be detected and a single error to be corrected. Since most soft error events are single-bit errors, EDAC/ECC protection will provide a significant reduction in soft failure rates (typically $> 10{,}000\times$ reduction in effective error rates) but at a higher cost in terms of design complexity, the additional memory required, and the inherent latency

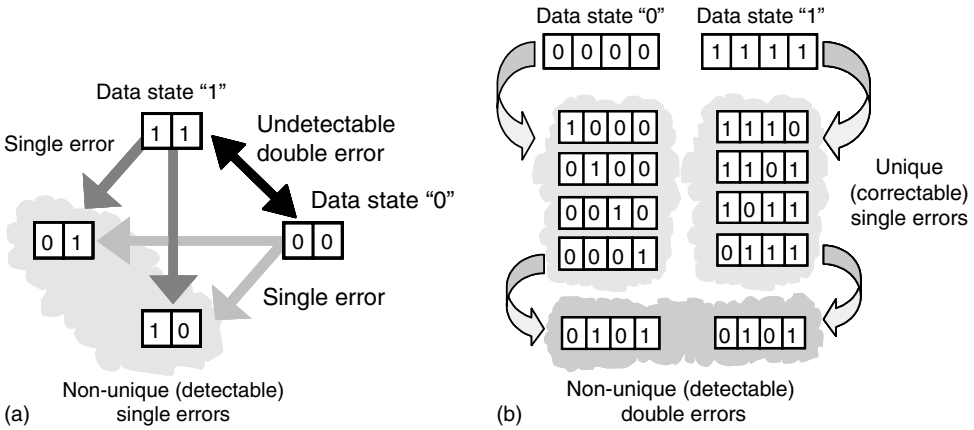


FIGURE 31.14 (a) Simplified representation of parity encoding/decoding. An additional bit is added to the data so that if a single error occurs the encoded word will be distinguishable from uncorrupted data and the error will be detectable. (b) Simplified representation of error correction encoding. Additional bits are used to uniquely identify the original data even after a single error has occurred—thus the detected error can also be corrected. Note that with both these methods multiple errors in a single data word can cause problems.

introduced during access, parity check, and correction. To ensure low failure rates, the memory design must also account for molecular beam epitaxies (MBEs) that can span 2 or more physical bits. In error corrected memories, it is recommended that the minimum row or column spacing between bits in the same logical "correction word" be at least 4 or 8 bits. The worst-case design would be a high-density memory with adjacent bits in the same correction word. In this type of layout, the efficacy of ECC would be limited by the MBE rate which, although a fraction of the intrinsic SER, would be orders of magnitude higher than a properly designed memory with ECC.

The analog of error correction in sequential logic involves the use of redundancy. There are fundamentally three ways to achieve this redundancy: create extra nodes within the logic elements so that data states are stored in more than a single location within the cell, create three or more identical logic elements feeding into a robust majority voting circuit so that an SEU in any one of the logic elements will be mitigated by the majority of good inputs from the other logic elements, and finally, the use of temporal sampling of logic inputs to filter out transients caused by SET and SEU. It should be noted that some of these methods can be combined to achieve high levels of reliability.

A robust device can be formed if multiple transistors associated with a particular data state are physically separated within the device to ensure that a "typical" single event cannot effect both transistors [42,43] (this approach is based on the fact that the probability of having multiple events in the same device node at the same time is exceedingly small). Known as state redundancy hardening, these types of designs can double the real-estate required but with proper design stand-by power and switching speed are not significantly impacted while SEU resistance is greatly increased. Multi-bit-upset events can still potentially upset these cells but optimizing the layout distances can also eliminate this risk—albeit at the expense of greater chip area.

The use of multiple identical logic paths with each path feeding into a hardened majority voting (two out of three) circuit allows a soft error in a single logic path to be ignored since the other two are the majority and, thus, the correct data "wins" the vote. This method uses more than three times the chip area and requires specialized simulation tools to identify the critical logic paths. Because of the high cost, one wants to protect only the most sensitive paths.

Time-multiplexed designs can also offer robustness against SEU and SET since the input is sampled several times before a decision on the output is made and thus input transients are not propagated. At an

increased cost, even more robustness can be built-in if time and spatial multiplexed techniques are combined [44,45]. However, in general the performance, power, and chip area penalty of these methods are not viable for mass-market commercial applications at this point.

The final and most ambitious form of redundancy is the use of duplicate or multiple redundant systems—where multiple, identical components are run in lock-step (executing the same code at the same time). In a dual-component system, a restart is issued when a mismatch between the processors is detected—a dual processor system is equivalent to parity protection in that errors can be detected but not automatically corrected. When more than two processor units are used, a majority voting strategy can be utilized so that restarting is not necessary. In other words with more than two processors running in lock-step, the system is self-correcting and avoids the latency associated with a system restart or instruction reloading. This is the most expensive redundancy scheme, but it does reduce soft failure rates to near-zero levels, providing the necessary reliability for some long-term remote or mission-critical applications.

31.6 Summary

Ionization collected from terrestrial radiation events can cause data errors leading to failures in electronic devices. At terrestrial altitudes three mechanisms are responsible for soft errors: the reaction of high-energy cosmic neutrons with silicon and other device materials, alpha particles emitted from trace radioactive impurities in the device materials, and the reaction of low-energy cosmic neutrons with high concentrations of ^{10}B in the device. The soft error sensitivity as a function of technology scaling for various memory and logic devices used to create advanced commercial microelectronic components revealed that while the soft error susceptibility of DRAM in a system is relatively unchanged by scaling, SRAM, and peripheral logic system susceptibility to soft errors is increasing rapidly with each new technology node. The efficacy of various methods to mitigate soft errors has been reviewed with the insight that memory system soft error reliability is best addressed by employing EDAC techniques, while sequential logic robustness can best be improved by design hardening. In closing, the reader is reminded that unlike other hard reliability mechanisms, where a single failure criteria can generally be applied to all products built with a particular technology, the actual customer impact of soft errors is extremely application-dependent. For single-user commercial applications, soft errors are typically not a concern, while for larger (multi-chip) or high-reliability applications, full error correction and/or redundancy techniques are mandatory.

References

1. Dodd, P. E., and L. W. Massengill. "Basic Mechanisms and Modeling of Single-Event Upset in Digital Microelectronics." *IEEE Trans. Nucl. Sci.* 50 (2003): 583–02.
2. Sexton, F. W. "Destructive Single-Event Effects in Semiconductor Devices and ICs." *IEEE Trans. Nucl. Sci.* 50 (2003): 603–21.
3. Satoh, S., Y. Tosaka, and S. A. Wender. "Geometric Effect of Multiple-Bit Soft Errors Induced by Cosmic Ray Neutrons on DRAM's." *Electron Device Lett., IEEE* 21 (2000): 310–12.
4. Maiz, J., S. Hareland, K. Zhang, and P. Armstrong. "Characterization of Multi-Bit Soft Error Events in Advanced SRAMs." *IEDM Tech. Dig.* (2003): 2141–44.
5. Koga, R., S. H. Penzin, K. B. Crawford, and W. R. Crain. "Single Event Functional Interrupt (SEFI) Sensitivity in Microcircuits." In *Proceedings of 4th RADECS*, 311–8, September 1997.
6. Benedetto, J., P. Eaton, K. Avery, D. Mavis, M. Gadlage, T. Turflinger, P. Dodd, and G. Vizkelethyd. "Heavy Ion-Induced Digital Single-Event Transients in Deep Submicron Processes." *IEEE Trans. Nucl. Sci.* 51 (2004): 3480–5.
7. Bruguier, G., and J-M. Palau. "Single Particle-Induced Latchup." *IEEE Trans. Nucl. Sci.* 43 (1996): 522–32.

8. Dodd, P. E., M. R. Shaneyfelt, J. R. Schwank, and G. L. Hash. "Neutron-Induced Latchup in SRAMs at Ground Level." In *Proceedings of 41st Annual IEEE International Reliability Physics Symposium (IRPS), EDS*, 51–5, April 2003.
9. Ziegler, J. F., and J. P. Biersack. The *Stopping and Range of Ions in Matter*, [version SRIM-2006.06 (c) 2006].
10. Ziegler, J. F., and W. A. Lanford. "The Effect of Sea Level Cosmic Rays on Electronic Devices." *J. Appl. Phys.* 52 (1981): 4305–18.
11. Normand, E. "Single Event Effects in Avionics." *IEEE Trans. Nucl. Sci.* 43, no. 2 (1996): 463.
12. Moscow Neutron Monitor web site: http://helios.izmiran.rssi.ru/cosray/main.htm (accessed on January, 2007).
13. Goldhagen, P. "Cosmic-Ray Neutrons on the Ground and in the Atmosphere." *Mater. Res. Soc. Bull.* 28 (2003): 131–5.
14. Gordon, M. S., P. Goldhagen, K. P. Rodbell, T. H. Zabel, H. H. K. Tang, J. M. Clem, and P. Bailey. "Measurement of the Flux and Energy Spectrum of Cosmic-Ray Neutrons." *IEEE Trans. Nucl. Sci.* 51 (2004): 3427–34.
15. Wrobel, F., J. M. Palau, M. C. Calvet, O. Bersillon, and H. Duarte. "Incidence of Multi-Particle Events on Soft Error Rates Caused by n-Si Nuclear Reactions." *IEEE Trans. Nucl. Sci.* 47 (2000): 2580–5.
16. Wrobel, F., J. M. Palau, M. C. Calvet, and P. Iacconi. "Contribution of SiO_2 in Neutron-Induced SEU in SRAMs." *IEEE Trans. Nucl. Sci.* 50 (2003): 2055–9.
17. May, T. C., and M. H. Woods. "A New Physical Mechanism for Soft Error in Dynamic Memories." In *Proceedings of 16th Annual International Reliability Physics Symposium (IRPS), EDS*, 33–40, 1978.
18. Low-alpha lead symposium proceeding, Lawrence Livermore National Laboratory, IPC-TM-650, Number 2.3.44, February 26, 1997.
19. Fleischer, R. L. "Cosmic Ray Interactions with Boron: A Possible Source of Soft Errors." *IEEE Trans. Nucl. Sci.* NS-30, no. 5 (1983): 4013–15.
20. Oldham, T. R., S. Murrill, and C. T. Self, "Single Event Upset of VLSI Memory Circuits Induced by Thermal Neutrons." *HEART Conference*, 1986.
21. Baumann, R. C., T. Z. Hossain, S. Murata, and H. Kitagawa. "Boron Compounds as a Dominant Source of Alpha Particles in Semiconductor Devices." In *Proceedings of 33rd Annual International Reliability Physics Symposium (IRPS), IEEE EDS*, 297–302, 1995.
22. Baumann, R. C., and E. B. Smith. "Neutron-Induced ^{10}B Fission as a Major Source of Soft Errors in High Density SRAMs." *Elsevier Microelectron Reliability.* 41, no. 2 (2001): 211–8.
23. Hsieh, C. M., P. C. Murley, and R. O'Brien. "A Field-Funneling Effect on the Collection of Alpha-Particle-Generated Carriers in Silicon Devices." *IEEE Trans. Electron Device Lett.* 2 (1981): 686–93.
24. Dodd, P. E., and F. W. Sexton. "Critical Charge Concepts for CMOS SRAMs." *IEEE Trans. Nucl. Sci.* 42 (1996): 1764–71.
25. Massengill, L. W. "Cosmic and Terrestrial Single-Event Radiation Effects in Dynamic Random Access Memories." *IEEE Trans. Nucl. Sci.* 43 (1996): 576–93.
26. Buchner, S., M. Baze, D. Brown, D. McMorrow, and J. Melinger. "Comparison of Error Rates in Combinational and Sequential Logic." *IEEE Trans. Nucl. Sci.* 44 (1997): 2209–16.
27. Hareland, S., J. Maiz, M. Alavi, K. Mistry, S. Walstra, and C. Dai. "Impact of CMOS Process and Scaling and SOI on Soft Error Rates of Logic Processors." In *Proceedings of IEEE International Symposium on VLSI Technology*, 73–4, 2001.
28. Baumann, R. "The Impact of Technology Scaling on Soft Error Rate Performance and Limits to the Efficacy of Error Correction." *IEEE IEDM Tech. Dig.* (2002): 329–32.
29. Zhu, X., R. Baumann, C. Pilch, J. Zhou, J. Jones, C. Cirba. "Comparison of Product Failure Rate to the Component Soft Error Rates in a Multi-Core Digital Signal Processor." In *Proceedings of 43rd Annual IEEE International Reliability Physics Symposium (IRPS)*, 209–14, April 2005.
30. Gadlage, M. J., R. D. Schrimpf, J. M. Benedetto, P. H. Eaton, T. L. Turflinger. "Modeling and Verification of Single Event Transients in Deep Submicron Technologies." In *Proceedings of 42nd Annual International Reliability Physics Symposium (IRPS), IEEE EDS*, 673–4, April 2004.

31. Dodd, P. E., M. R. Shaneyfelt, J. A. Felix, and J. R. Schwank. "Production and Propagation of Single-Event Transients in High-Speed Digital Logic ICs." *IEEE Trans. Nucl. Sci.* 51, no. 6 (2004): 3278–84.
32. Baumann, R., and T. Hossain. Electronic device and process achieving a reduction in alpha particle emissions from boron-based compounds essentially free of boron-10. United States Patent 5,395,783, March 7, 1995.
33. Hwang, M., R. McKee, and R. Baumann. Thermal neutron shielded integrated circuits. United States Patent 6,239,479, May 29, 2001.
34. Xu, Y. Z., H. Puchner, A. Chatila, O. Pohland, B. Bruggeman, B. Jin, D. Radaelli, S. Daniel. "Process Impact on SRAM Alpha-Particle SEU Performance." In *Proceedings of 42nd Annual IEEE International Reliability Physics Symposium (IRPS)*, 294–9, April 2004.
35. Musseau, O. "Single-Event Effects in SOI Technologies and Devices." *IEEE Trans. Nucl. Sci.* 43 (1996): 603–13.
36. Roche, P., G. Gasiot, K. Forbes, V. O'Sullivan, and V. Ferlet. "Comparisons of Soft Error Rate for SRAMs in Commercial SOI and Bulk below the 130-nm Technology Node." *IEEE Trans. Nucl. Sci.* 50 (2003): 2046–54.
37. Cannon, E. H., D. D. Reinhardt, M. S. Gordon, and P. S. Makowenskyj. "SRAM SER in 90, 130, and 180 nm Bulk and SOI Technologies." In *Proceedings of 42nd Annual International Reliability Physics Symposium (IRPS), IEEE EDS*, (2004): 300–4.
38. Roche, P., G. Gasiot, K. Forbes, V. O'Sullivan, and V. Ferlet. "Comparisons of Soft Error Rate for SRAMs in Commercial SOI and Bulk below the 130-nm Technology Node." *IEEE Trans. Nucl. Sci.* 50, no. 6 (2003): 2046–54 (Part 1).
39. Nguyen, D. N., S. M. Guertin, G. M. Swift, and A. H. Johnston. "Radiation Effects on Advanced Flash Memories." *IEEE Trans. Nucl. Sci.* 46, no. 6 (1999): 1744–50.
40. Benedetto, J. M., W. M. De Lancey, T. R. Oldham, J. M. McGarrity, C. W. Tipton, M. Brassington, and D. E. Fisch. "Radiation Evaluation of Commercial Ferroelectric Nonvolatile Memories." *IEEE Trans. Nucl. Sci.* 38, no. 6 (1991): 1410–4 (Part 1).
41. Rockett, R. "Simulated SEU Hardened Scaled CMOS SRAM Cell Design Using Gated Resistors." *IEEE Trans. Nucl. Sci.* 39 (1992): 1532–41.
42. Bessot, D., and R. Velazco. "Design of SEU-Hardened CMOS Memory Cells: The HIT Cell." In *Proceedings of 2nd RADECS*, 563–70, September 1993.
43. Calin, T., et al. "Topology-Related Upset Mechanisms in Design Hardened Storage Cells." In *Proceedings of 4th RADECS*, 484–8, September 1997.
44. Anghel, L., and M. Nicolaidis. "Cost Reduction and Evaluation of a Temporary Faults Detecting Technique." In *Proceedings of Design, Automation and Test European Conference 2000*, 591–8, March 2000.
45. Mavis, D. G., and P. H. Eaton. "Soft Error Rate Mitigation Techniques for Modern Microcircuits." In *Proceedings of 40th Annual International Reliability Physics Symposium (IRPS), IEEE EDS*, 216–25, April 2002.
46. Private communication with Xiaowei Drng.
47. Private communication with Xiaowei Zhu.

32

Integrated-Circuit Packaging

Michael Lamson
Texas Instruments, Inc.

Andreas Cangellaris
The University of Illinois

Erdogan Madenci
University of Arizona

32.1 Introduction **32**-1
32.2 Electronic Packaging Challenges **32**-2
Transition of Packaging Formats—Traditional Packages—Area Array Interconnect • System in a Package—3D Packaging • Roadmap Projections • The Challenge of High Frequency Electrical Performance • The Challenge of Interconnect and Mechanical Reliability • The Challenge of Thermal Management • Research Challenges
32.3 Electrical Modeling and Behavior of Packages..... **32**-9
Electrical Parameters • Overview of Modeling Methodologies
32.4 Hygro-Thermo-Mechanical Behavior of Packages..... **32**-13
Solder Joint Fatigue Failure • Hygro-Thermo-Mechanical Analysis • Thermal Deformation of Flip-Chip Packages • Characterization of Material Properties • Modeling Verification for Flip-Chip Packages • Summary of Thermo-Mechanical Behavior
References..... **32**-27

32.1 Introduction

The role of packaging in semiconductor electronic applications is to protect and preserve the performance of the semiconductor device from electrical, hygro-thermo-mechanical, and chemical corruption or impairment. This role has become more and more important as well as difficult to execute as device performance, complexity, and functionality increase with each succeeding generation of technology. These generations are succinctly captured in the 2005 International Technology Roadmap for Semiconductors (ITRS) [1], published by the Semiconductor Industry Association, and they have been highlighted in the Overall Roadmap Technology Characteristics section of the ITRS. The present article will not focus on many types of packaging that have been developed in the past or seek to give a fundamental treatment of the complex phenomena that occur in the conventional package during use. These topics have been covered in a comprehensive manner in other sources [2,3]. Instead, the authors will try to acquaint the semiconductor engineer with the areas of fundamental importance in packaging; as it evolves from present practices into the future and provide some overview of current modeling and analysis methods being used by package design engineers.

Current and projected packaging solutions will be discussed in terms of their electrical and hygro-thermo-mechanical behavior as to how these factors affect the performance of electronic devices. From the projected device characteristics are derived the needs for packaging and assembly technologies that

will support and enable the dynamic growth of the semiconductor industry. The most apparent changes in chips have been driven by the increase in chip speed, its size, the transistor density, and power dissipation. The packaging solutions for these chips have been driven by the need for greater and greater interconnect to and from the chip, the need for more effective management of chip and system mixed signal and high frequency behavior, and high capacity thermal management along with the continued reduction in the power supply voltage levels. The chip to package interconnections need to be capable of operating at a very high speed as well as carry larger currents than have been customarily associated with integrated circuits in the past.

In addition to these more or less linear projections of the devices, the interconnect on the chip itself has undergone profound changes that has been driven by the need for enhanced chip performance. This chip level interconnect has not only become more dense per unit area, but the materials set has been changed entirely [4]. This change in chip level materials is the most thoroughgoing modification of interconnect in 35 years, and, in turn, has generated the need for different materials for chip-to-substrate interconnect.

In some cases, specifically the envelope of packaging is shrinking into a minimalist version of the products that traditionally have been called packaging. The "chip scale package" has become established in radio frequency application in hand-held and portable devices [5,6]. It is in effect an "intrinsic package." An extrinsic package, which can be qualified as an entity separate from the chip, is being replaced by a packaging solution that incorporates the chip as an integral mechanical part of the system. Thus, the chip and its substrate have become a packaging system that must protect the die from moisture and contaminants in the environment, preserve its electrical performance, and manage the thermal load the system dissipates. In addition, to this change, there are non-technical challenges that will limit the solutions available to the packaging engineer. One of the more significant challenges is the issue of the cost of packaging. In some cases, many of the needs of high performance systems can be met with some version of high density interconnect or multichip module packaging solutions that have been developed over the last 10–15 years. The cost and availability of these solutions, however, have generally not met the targets for manufacturability and cost desired by the users. Their adoption in commercial applications has therefore been slow. Additional invention and development will be needed, as affordable packaging solutions are deployed to support the longer term needs shown in the 2005 ITRS. All of these capabilities of future packaging solutions will, of course, require a more advanced and integrated system of software to design, model, and to predict electrical performance and the hygro-thermo-mechanical reliability than is presently available.

32.2 Electronic Packaging Challenges

The need for better packaging design tools and processes can be most usefully discussed in terms of the needs of electronic systems for increased functionality, higher performance, and reliable behavior. These needs are discussed in terms of challenges that must be met. In this introduction the ITRS defined needs and challenges will be outlined. In the following two sections will be discussed the hygro-thermo-mechanical tools for measurement and control, and the software systems for electrical design and modeling the packaging solutions needed in the future.

32.2.1 Transition of Packaging Formats—Traditional Packages—Area Array Interconnect

The driving force for the development of new packaging solutions is the development of new and higher performing electronic devices. Such devices support the pervasive applications of information technology (transmitters and receivers, wireless chip sets, high speed modems, computation (microprocessors, memory)), and milieu control (automobile, home climate, shop floor management, and aircraft avionics), and array of hand-held devices that enhance the lives of the citizens of modern society.

The traditional packaging technology that has supported device protection and performance requirements has involved such structures as the dual in-line package (DIP) and the quad flat package (QFP) [2]. These formats continue to support vast numbers of commercial semiconductor products. In the majority of applications for these technologies, the chip is electrically connected to the package via wire bonding [5]. The DIP uses pins to connect the packaged chip to the electronic system and these pins are inserted into a printed circuit board (PCB) or socket during assembly. The leads of the QFP, on the other hand, are solder mounted onto the surface of the PCB rather than being inserted into the board as is the DIP. This surface mount technology (SMT) can support many more packages to board leads than the DIP in a much smaller "footprint" on the board. At about 250 leads, however, the increasing difficulty in manufacturing the QFP format has tend to establish a practical limit to its further extension to higher lead count. Accordingly, the commercial industry has moved towards the ball grid array (BGA) format to support higher lead count packages. In Figure 32.1a the wire bonded BGA is illustrated. This format tends to be limited by the peripheral chip pads that can be supported by wire bonding. As the number of transistors in a chip increases, so must the number of chip pads needed to support them with power, ground, clock, and signal. Thus, as chip complexity increases, the practical limit for peripheral chip pads tends to be exceeded. Wire bonding technology has been extremely successful in supplying chip-to-substrate interconnect at affordable costs. The peripheral wire bonding for integrated circuit connections will continue the bonding pitch reduction and cost improvements that we have seen in the past [6]. However, it is limited in its pad count to about 900 bond pads on the chip. At about 900 chip pads, the production problems associated with wire bonding begin to affect product throughput as well as yield. Ultimately these problems equate to increased cost. In a parallel vein, wire bonding is more applicable to peripherally bonded chips. With peripheral I/O connections, the voltage drop along the chip's power and ground lines will impact the signal/noise immunity for medium and high power chips. When circuits in the chip interior are connected to those near the periphery, they will be affected by this voltage instability [7,8]. For these reasons, it is necessary to adopt an area array configuration for high lead count chip-to-package interconnections [9] in the cost-performance and high performance market segments. Area array interconnect allows cost effective interconnect beyond the practical limit for wire bonding. The flip chip BGA, illustrated in Figure 32.1b, has been well accepted as an important vehicle for providing area array interconnect.

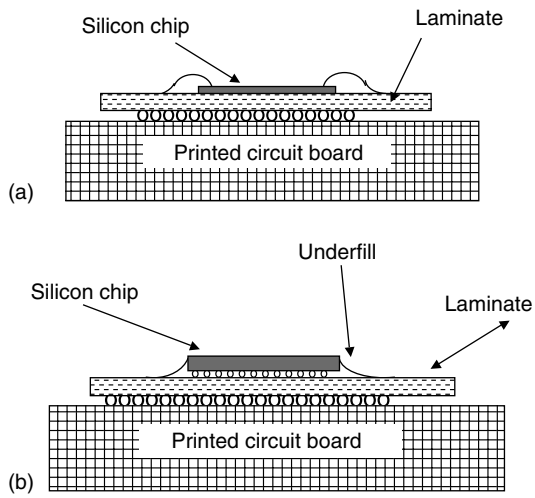


FIGURE 32.1 (a) Wire-bonded ball grid array (BGA); (b) flip-chip die assembled on BGA.

32.2.2 System in a Package—3D Packaging

Combining the attributes of BGA formats and expanding the concept vertically with multiple chips has spawned the more recent concept of system in a package (SIP). In this concept, multiple chips are assembled in a vertical mode, either in separate packages such as package on package assemblies or in terms of multiple stacked dies in a package. The advantages here are improved board space, integrated functionality in close proximity, possible add on technologies, improved time to market, and lower cost. These 3D formats can utilize both flip chip and wire bond assembly methodologies. The process of chip-package co-design becomes a very important concept in designing these devices. Again, efficient computer modeling, simulation, and design tools are very important for achieving a functional device. Obvious issues are also thermal management and die thinning; handling methods are important and these are noted as difficult challenges in the 2005 ITRS.

32.2.3 Roadmap Projections

The Roadmap projects a need for steadily increasing transistor density to supply more and more functional capability from the same area of silicon. For 40 years we have seen the success of this strategy of designing larger and larger integrated circuits. The cost to produce the electronic functions has steadily decreased by about 25% per year per function. In step with this cost reduction we have seen system reliability increase as more and more functionality is swept onto the chip and off the PCB. As the transistors have increased in number, they have been reduced in size to support higher device density and progressively increased chip speed. The on-chip development that has driven this migration is traced in Table 32.1 which shows some of the assembly and packaging requirements for hand-held or low cost and cost-performance market segments [1].

32.2.4 The Challenge of High Frequency Electrical Performance

The increasing use of higher and higher frequencies in commercial electronics is given of in today's computer and communication systems. In Table 32.1 we see that we are well into the realm of gigahertz operation. We also see in today's professional journals and the trade papers that many trends in computing and telecommunication are being advanced that will bring ever higher levels of speed to commonly available electronics.

The issues that enhanced device performance bring to packaging solutions; involve the preservation of signal integrity and timing as the signal moves within a chip and from chip to chip within the system. The off-chip interconnect must preserve chip performance, which will imply controlled transmission characteristics, electromagnetic interference (EMI) shielded interconnect, and power integrity control. All this must be achieved at an affordable cost. In addition, due to signal loss in the interconnect, some applications will require the chips and substrates to incorporate passives to support noise control, and the integrity of ground and supply voltages [10]. Many of these individual capabilities are currently commercially available (such as low loss Teflon substrates and embedded passives in ceramic substrates [11]), but their costs continue to restrict their use to the high performance, the military, and the aerospace electronics market. Cost reduction is needed, as these capabilities become available to support cost-performance and commercial applications.

One of the most important of these new capabilities is the software for the design and optimization of the new packages required for future devices. These are the tools that will enable the effective and timely design for "first time success" of new packages. It is these packages that are becoming more and more complex and demanding, and that are needed to maintain the Roadmap momentum in the future. The tools for the modeling and simulation of electrical performance need to be enhanced to accommodate the high frequency, high power digital, and mixed signal performance [12]. Better tools for the accurate and efficient estimation of electrical parasitics, cross talk, power and ground noise, and EMI continue to be needed. They need to comprehend the impact of high-speed performance on system layout and noise management [13]. Their application to the total electronic system—chips, interconnect, and

TABLE 32.1 Selected Performance Characteristics of Packaged Chips from the 2005 International Technology Roadmap for Semiconductors (ITRS): Microprocessor Product Needs

Year of First Product Shipment (year 1)	2005	2006	2007	2008	2009	2010	2012
Feature size (nm)	90	70	65	57	50	45	36
Logic xstors (M/cm^2) (packed, incl. SRAM)	174	219	276	348	438	552	876
Chip size (mm^2)							
Cost performance	140	140	140	140	140	140	140
High performance	600	630	662	695	729	755	750
Package cost (cents/pin)							
Hand-held[a]	0.27–0.50	0.26–0.49	0.25–0.48	0.24–0.47	0.23–0.46	0.22–0.45	0.21–0.42
Cost-performance[b]	0.68–1.17	0.66–1.11	0.64–1.05	0.63–1.00	0.62–0.96	0.61–0.94	0.58–0.90
Number of package pins (balls)							
Hand-held	134–550	140–578	148–606	150–636	160–668	170–700	188–774
Cost-performance	550–900	550–900	600–1,088	600–1,198	660–1,318	660–1,450	720–1,754
Power density (W/mm^2)							
Hand-held	2.80	3.00	3.00	3.00	3.00	3.00	3.00
Cost-performance	0.65	0.70	0.74	0.79	0.83	0.85	0.89
Minimum logic V_{dd} (V)							
Cost-performance	1.0	0.9	0.9	0.8	0.8	0.6	0.6
Across chip clock (MHz)[c]	5,204	6,783	9,285	10,972	12,369	15,079	20,065
Chip-to-board clock (MHz)							
High Performance	3,125	3,906	4,883	6,103	7,629	9,536	14,900

[a] Hand-held: battery-powered products.
[b] Cost performance: notebook computers, desktop PC's and telecommunications.
[c] High performance: high end workstations, servers, avionics, and the most demanding applications.

substrates will be needed. These tools must model the chip as a part of a packaging solution to produce new electronic systems in the ever-shortening time to market our industry needs.

32.2.5 The Challenge of Interconnect and Mechanical Reliability

We see in Table 32.1 that the transistor density (millions of transistors/square cm) increases from 174 in 2005 to 876 in 2012 for static random access memories (SRAMs). These transistors must be supported with power, ground, and signal I/O. For a chip with a given circuit or transistor count, the lead count corresponds to an application of Rent's Rule [14], which relates the two quantities. In the Roadmap, typically the exponential coefficient used is about 0.3.

In past Roadmaps [15], packaging/assembly technology has not scaled in a manner that parallels and accommodates the scaling of on-chip technologies. The 1997 National Technology Roadmap for Semiconductors reflects a need for chip-to-substrate and package-to-board interconnect technologies to provide a scaling that parallels the on-chip interconnect technology. The most significant challenges that this poses for packaging solutions lie in accommodating the higher chip interconnect density with appropriate chip-to-substrate interconnect. The technology that accomplishes this goal of scaling is, of course, area array interconnect. This technology has been in applications for about 30 years within companies such as IBM and Delco Electronics. It is now used throughout the semiconductor industry as an accepted, reliable way for packaging cost performance and high performance devices. The area array format is illustrated in Figure 32.2 as a flip-chip BGA. This adaptation of flip-chip to plastic packaging allows low cost support for very high levels of on-chip interconnects. This format has been the source of significant technology challenges.

In Figure 32.2 the chip is shown as being flip-chip assembled onto a high-density substrate. The high-density substrate uses a conventional PCB as its core, but the fan out or escape from the chip is achieved using surface layer circuitry (SLC)—which is also referred to as a built-up layer. This SLC is patterned using spin on materials such as polyimide for the inter layer dielectric (ILD) and electroplated copper interconnect. The vias are usually photo-defined and as such are of much finer dimensions than the plated through hole used in the PCB. There are balanced layers of SLC placed on either side of the PCB to support a balanced stress between the two levels of interconnect. This is primarily done to improve the flatness of the substrate. Needless to say this added processing adds significantly to the cost of the substrate.

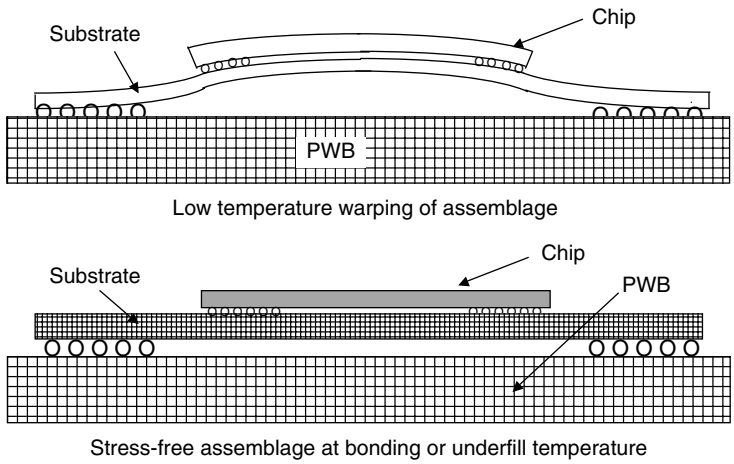


FIGURE 32.2 The flexing of chip/laminate ball grid array (BGA) on printed wiring board (PWB). Underfill reduces transverse stress by creating a normal stress, which is borne mostly by the underfill polymer the chip and the laminate substrate, not the solder.

32.2.5.1 Chip-Level Interconnect Modifications

The development of packaging solutions, while complex in itself, is further complicated by the profound changes seen in the wafer level interconnect materials. These materials have migrated from the historic Al/SiO_2 to Cu/low-*k* [1] for the conductor and ILD materials, respectively. The different wafer interconnects metallurgy require assembly metals (such as the under bump metallization) that can accommodate both Al and Cu on the wafer bond pads. The new ILD's are made of low dielectric constant (2.0–3.0) material. Many of these materials have low mechanical strength and low thermal conductivity compared to SiO_2 as the ILD [1]. These differences mean that the wafer will be a weaker mechanical component in the packaging. The on-wafer interconnect comprises multiple lines of ultra-fine metal interconnect embedded in the low-*k* dielectric. These structures support little strain and can become a significant reliability risk. Mitigating this strain requires packaging solutions that impose little stress on the chip during thermal cycling. It is noted in the 2005 ITRS that even now the impact of mold compounds on the Cu/low-*k* dielectrics are not well understood.

32.2.5.2 Underfill Polymers in Flip-Chip Processing

The use of underfill has been a significant factor in making flip chip onto laminates, a viable technology today [16]. The coefficient of thermal expansion (CTE) of silicon is about 2.6 ppm/°C, and that of the glass fiber reinforced printed wiring board (PWB) is about 17 ppm/°C. The BGA substrate is the intermediary between the silicon and the PCB with a CTE of 6 ppm/°C for ceramic BGA, CBGA and 17 ppm/°C for organic BGA. The large CTE difference between CBGA and organic PWB limits the CBGA body size to about 32 mm. On the other hand, the large CTE difference between plastic BGA (PBGA) and the silicon chip puts a significant shear stress on the flip chip interconnect and almost any commercial application is impractical without the use of underfill. Underfill encapsulation between the front-side of the chip and the top side of the PBGA substrate distributes the stress over the entire chip surface, and so reduces the stress on the flip chip solder joints. The impact of this CTE mismatch is that the entire assembly flexes during thermal cycling to distribute stress away from the solder balls. This behavior is illustrated in Figure 32.3 and it will be treated in full in Section 32.3.

It is important that the underfill material has good mechanical integrity and interfacial adhesion to both the chip and the substrate surfaces and the vertical direction CTE of the underfill is compatible with that of the solder joints. The cost associated with the dispensing of the underfill into the thin gap between the chip and the substrate and the curing time are other challenges. Note that as the gap is reduced there will be some height of this gap where "conventional" underfills will be very difficult to inject between the chip and the substrate. At that point a "no-flow" underfill will be required [17], or a whole new paradigm will be required. Indeed, as soon as a practical no-flow underfill is available, its processing costs should be less expensive than a flowed material.

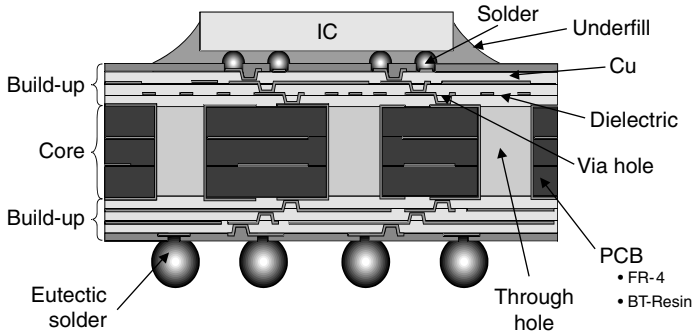


FIGURE 32.3 Flip-chip ball-grid array package.

32.2.5.3 Alternate Solutions

The application of underfill has created a chip that is bonded to the substrate and hence has become a part of the overall reinforced structure [18]. This redistributes the thermo-mechanical strain away from the solder balls into the system as a whole—which makes the flip-chip PBGA type of assembly commercially viable. As was noted earlier, however, this bonding puts the chip in a strained state that is very different from that experienced in wire bonded applications [19], which may involve a reliability risk to the chip level interconnect.

There are several paths to mitigate this risk. One path is the use of chips of limited size in these high performance applications. A second path is in the development of alternate materials for substrates that match more closely to the wafer in their CTE. Ceramic substrates answer this need at present, but cost is a significant issue in this case, and the compatibility with the PCB has become a critical problem. Even this alternate substrate, however, will not completely mitigate the stresses associated with power-on current surges that create critical thermal stresses in large chips. A third approach for chips above some size and power would be the use of an alternate compliant interconnects for chip to substrate connections. This interconnect is needed to accommodate the strain and shield the chip from these stresses. Also, as the solder balls become smaller, their compliance is reduced and they are more subject to strain and fatigue failure. Some mitigation of the strain is needed. A compliant interconnect, such as a contacting system based on wires or columns or compliant, conductive polymers may be needed as an alternate interconnect to mitigate this strain. This type of interconnect would be appropriate for large and high performance chips needing very high chip-to-package interconnect. A fourth path would be to utilize a nonresistive radiative interconnect such as optical chip to chip interconnect to avoid the thermal strain altogether.

32.2.6 The Challenge of Thermal Management

The high performance chips that will dissipate a heat load approaching 200 W in 2012, will require more actively managed thermal systems than those used commercially today. Again, a plethora of thermal management solutions in practice for high performance, high wattage systems are available today. Fluid flow heat exchangers can support very large loads. In addition, diamond heat spreaders and diamond micro-channel heat exchangers can be used to support thermal loads, an order of larger magnitude [20]. The issue in both cases is affordability. In portable laptop computers, this evolution has occurred relative to the introduction of heat pipes to manage the thermal load. A similar evolution will be needed to supply the solutions for thermal management in 2012.

32.2.6.1 Heat-Sink Attachment for Large and High Power Chips

We have seen that the CTE difference between the chip and the PBGA substrate will cause both to bend when the encapsulated chip/package entity is cooled from the curing temperature to room temperature [21]. The material and properties of the chip, the substrate, and the underfill will determine the curvature of this structure. The shear stress on the solder joints is reduced due to the bending. However, for a large chip, the back surface of the chip will have a corresponding warpage, which will likely be greater than the thickness of thermal grease used to connect a heat sink to the package. During the on–off power cycles in daily operations, the packaged chip may push out the thermal grease and degrade the thermal performance. The use of thermal adhesive in attaching a rigid heat sink to the backside of the chip will prevent the bending of the chip, and the shear stress to the solder joint and encapsulant may remain excessively large. These are some of the factors that must be considered in developing a thermal solution for chip packaging.

An alternative to thermal grease or adhesive can be the use of a cooling piston, similar to that used in the thermal conduction module [22], on the backside of a large chip. The chip may still flex as was discussed above, but the piston can move to adjust to the movement. Finally, flexible thermal heat sinks or non-contacting thermal methods may present themselves as useful methods for thermal management

and control. Therefore, thermal management and the associated cost are also part of the future challenges.

32.2.7 Research Challenges

The packaging solutions that will support and enable the continuation of the semiconductor industry growth as well as achieve the needed cost reductions projected for 2012 will require significant invention beyond and enhancement of current capabilities. The advent of flip chip assembly has at one stroke solved the issues of achieving a very high chip to package interconnect, achieving a low interconnect impedance, and maintaining a small footprint for the chip on the board. The invention of underfill technology has enabled chip to laminate assembly, which promises to provide greater affordability to the large and dense chips of the future. We are finding that the packaging solutions for the future are much more complex than our current technologies. Adding to the complexity, the wafer is changing much more that it has since multilevel interconnect was introduced some 20 years ago. Resolving this complex situation will require R&D focused on understanding the mechanics, the thermal behavior, and the electrical performance of the new materials and chips that will be used in the packaging solutions of 2012. The focus of this effort will be to create software tools to design and to model these new configurations. The sheer magnitude of choices before U.S. means that experimental investigations of every choice are too expensive and time consuming to be practical in developing new solutions. There must be software tools to lead the development. These tools will be based on developing a fundamental understanding of the material behavior, the interfacial interactions, and the electrical, mechanical, and thermal performance of the constructs that will be the packaging solutions of 2012. Experimental test vehicles will be needed to validate the models and confirm the reliability of these constructs.

In addition, to the need for a fundamental understanding of the various aspects of present developmental packaging solutions, we have seen that there are areas where new and enhanced capabilities are needed. We have also seen that the present solutions leave a lot to be desired in terms of both capabilities and cost control. Further invention will be required to provide the enhanced solutions that will provide cost effective and high performance solutions to the packaging of the future.

32.3 Electrical Modeling and Behavior of Packages

The significant influence of the package on the electromagnetic behavior of the device has been well noted earlier in this chapter. Accurate prediction of the character of this behavior is now one of the most important tasks in the design of state-of-the-art packaged electronic systems. Indeed, the fast switching speeds, dense structures, high currents, and low supply voltages of current and future devices makes this an imperative in order to insure a successful device.

With respect to the electromagnetic performance, the major electrical roles of the package are to transmit signals properly between the integrated circuit chip(s) and the external circuits and to supply stable power to the device from the system power supply. In some cases, the package may also provide electromagnetic shielding between the chip and the external circuitry.

Proper transmission implies that signals are transferred through the signal distribution network (SDN) with an acceptable level of distortion and time delay. Without this control, logic errors may occur causing the ultimate failure of the system. There are many causes for distortion including attenuation of the signal voltage levels, reflections from conductor discontinuities, and crosstalk [23]. The management of these factors comes under the heading of "signal integrity."

Stable power means that the power and ground voltages are controlled to within acceptable variations from their dc levels. The package power and ground supply paths, or power distribution network (PDN), always exhibit some voltage drop due to the level ($\Delta V = IR$) and rate of change [$\Delta V = L(\mathrm{d}I/\mathrm{d}T)$] of the current being supplied. These fluctuations in ground and power levels will manifest themselves as variation in the signal level and reference potentials causing further distortions of the

output, and possibly, false switching [24]. This phenomenon is frequently referred to as "ground bounce" or "power droop." Further, the PDN can exhibit resonance issues causing uncontrolled variations in the reference voltages for the device and possibly radiation problems. The PDN may also be a source of coupled noise for signal nets that utilize the PDN as there reference power/ground system. Management of these effects is termed "power integrity."

These phenomena involving the SDN or PDN are not generally independent and may require a comprehensive analysis methodology to account for significant coupled elements to accurately predict the electrical behavior of the system under development.

Package electrical design is normally done by an iterative process involving successive time or frequency domain simulations with user-driven geometry changes between simulations, although there are examples of both closed-form design expressions (see, for example, [23]) and automatic optimizing software programs for design [25]. The simulations may be based on extraction of equivalent electrical parameters (resistance, L (inductance), and capacitance, RLC) of a packaging structure and subsequent simulation using a circuit simulator such as Simulation Program with Integrated Circuit Emphasis (SPICE) [26], or the simulations may be performed by full solution (usually numerical) of the four coupled Maxwell's equation (so-called full-wave solution) in either time [27] or frequency [28] domain. This second technique (i.e., full wave) is less commonly used because it requires very long computer run times. Alternatively, new approaches to efficient simulation of packaging structures involve combining simplified frequency-domain models of linear portions of a packaging network (see, for example, [29]) with a modified SPICE simulation program [30,53] to achieve rapid time-domain simulation of either lossless or frequency dependent lossy structures. These may involve transmission lines with frequency-dependent elements [30,45].

32.3.1 Electrical Parameters

For the case of using an equivalent circuit for simulating package electrical behavior, a computer model may generate electrical parameters in terms of resistance, capacitance, and inductance, based on the conductor geometry and material properties of the package structure. It is important to keep in mind that, although, package conductors exhibit resistive, capacitive, and inductive properties; they are not resistors, capacitors, or inductors in the sense of point circuit element. These properties must be thought of as being distributed along the length of the structure. Indeed, if the signal rise time, t_r, is fast enough (i.e., t_r is less than two times the transit time, t_t, of the signal, through the lead as a rule of thumb), the structure will behave as a transmission line and will require more than a single RLC element to model the electrical behavior [31]. At higher speeds, it will also be important to control a related parameter, the characteristic impedance, Z_0 of the transmission line structure to ensure maximum power transfer through the package. In the high frequency limit for a single line over a reference plane, $Z_0 = \sqrt{(L/C)}$.

32.3.1.1 Resistance

Resistance represents the property of a conductor to exhibit a voltage drop between its terminals when a current is flowing. This is due to the dissipation of energy in the form of heat within the conductor. The value of the voltage drop can be calculated from Ohm's law as $\Delta V = IR$.

The resistance of a conductor will depend on the material resistivity, ρ (see Table 32.2). For dc conditions the resistance, R is calculated as $R = \rho l/A$, where ρ, resistivity; l, length; and A, area of cross section.

Resistivity varies significantly between package types. Plastic packages may use copper conductors ($\rho = 1.67$ μohm-cm) or Alloy42 as a conductor material ($\rho = 38.8$ μohm-cm). Co-fired ceramic packages may use tungsten/glass composite materials for conductor traces, which may have an effective resistivity of 25 μohm-cm.

The effect of resistance on package electrical behavior is usually minor since the IR voltage drop is generally small when compared to other sources. At low to moderate frequencies, the major source of IR

TABLE 32.2 Resistivities (μΩ-cm)

Material	ρ
Copper	1.67
Silver	1.59
Gold	2.35
Aluminum	2.65
Nickel	6.84
Iron	9.71
Palladium	10.8
Tin	11

drop is from the bond wires of packages with lead frames. However, at higher frequencies, the resistance will increase because the current tends to flow non-uniformly through the conductor cross section. This is due to a phenomenon called the skin effect in which, the current tends to crowd towards the conductor surface at high frequencies [32]. The frequency-dependent resistance, which results is called the skin resistance and must be taken into account for accurate simulation of waveforms with very short risetimes [33].

32.3.1.2 Capacitance

When a potential, ΔV, is impressed between two conductors in an insulating medium, an electric charge, Q will accumulate on the positive conductor ($-Q$ on the negative conductor). The ratio of the charge to potential difference, $Q/\Delta V$, is defined as the capacitance, C. If the insulating medium is not a vacuum, polarization charge will also develop within the medium increasing the total charge and hence the capacitance of the system. The degree of total charge will depend on the relative dielectric constant, ε_r of the insulating medium. Table 32.3 shows relative dielectric constant values for typical packaging materials. Of course, these values will vary depending on the source and condition (i.e., moisture content, etc.) of the material.

Capacitance is important to package electrical behavior since it provides a path for electrical energy to be transferred between conductors (crosstalk), it acts as an energy sink as potential increases, which results in delay, and it may act as a source of energy by supplying charge to the circuit. It takes into account the action of the electric field and thereby in conjunction with the inductance of a structure gives rise to propagating electromagnetic waves, which carry signals in digital and analog systems. The capacitance of a conductor system depends on all of the conductor surfaces within the system. If a new surface is introduced, the capacitance between all of the original surfaces will change. In general, the capacitances of a system are described in matrix format, with self- and mutual capacitances as matrix elements.

For typical package conductor structures, the calculation of capacitance requires computer software except for very simple structures. A particularly useful format for extracting and describing the system capacitance is the Maxwell matrix [34], in which the elements of the matrix are coefficients, which are related to specific capacitances in the structure.

TABLE 32.3 Dielectric Constants (Relative) of Some Packaging Materials

Material	ε_r
Mold compound (type)	4.0
Polyimide	3.5
Alumina	9.8
Silica	3.7

32.3.1.3 Inductance

When current flows in a conductor, a magnetic field is created in the region around (external) and within the conductor (internal). The ratio of the total magnetic flux, Φ, to current, I, is defined as the inductance, L. Hence, $L = \Phi I$. Since, energy is required to establish the magnetic field, a voltage drop will occur across the conductor as the field builds. This voltage drop is related to the time rate of change of current as $\Delta V = L(dI/dT)$.

The inductance of a package will depend on the current distribution within the conductor system including any reference or power planes. Its effective value is a function of both the forward and return path of the current (i.e., the inductance of a single conductor alone has no real meaning). This presents difficulties in modeling inductance, when conducting planes are involved. Efficient methods of modeling inductance have been developed [35,36] and these are utilized in modern modeling tools [37,38,47].

The internal self-inductance of a conductor is only important at low frequencies when the current is distributed uniformly over the conductor cross-section and for magnetic materials. At frequencies of interest in digital electronics, the skin effect excludes much of the current from the interior of the conductor and the internal self-inductance will approach zero, leaving the external inductance as significant [39].

Inductance is important to package electrical design since it produces voltage drops in response to rapidly varying current levels in the circuit. In the ground and power paths in the package, for example, a sudden demand for current by the chip in response to signal switching will cause the voltage levels of the power system to change [40,41]. This can lead to simultaneous switching output noise or power supply collapse [23,24], either of, which may cause logic or timing errors and concomitant chip failure. Inductance is also important as a source of crosstalk between conductors, since the magnetic field due to current in one lead can induce voltage on a neighboring lead. This is related to the mutual inductance.

32.3.2 Overview of Modeling Methodologies

All computer based modeling approaches of interest to package and interconnect electrical analysis can be classified as integral equation or differential equation based [43]. This classification refers to the way that Maxwell's equation is solved. We will consider each method separately in order to compare their strengths and weaknesses as they pertain to package modeling. Various computer codes utilizing these methods are available commercially and from university sources [51,52,54].

32.3.2.1 Differential Equation Methods

The two most prevalent types of methods in the differential category are the finite element method (FEM) and the finite difference method. These methods utilize the discretization of Maxwell's equation in the problem space at hand. Due to the curl relationship between the electric and magnetic fields, the coupling is strongly localized; resulting in sparse matrices to be dealt with in the solution path for the finite difference method. Similarly, the FEMs also have the advantage of dealing with sparse matrices. In addition, the finite methods can model very complex geometries and materials properties with a high degree of accuracy given the use of modern mesh generation programs. In contrast, to the advantages noted, the finite methods require very large matrices to be solved, driving relatively long solution times.

For electromagnetic applications, it has been reported that the most popular time domain field solver is the finite difference time domain method [42]. The method is simple and effective utilizing a Cartesian grid to discretize the problem space of interest and solving for the electric and magnetic fields on alternate cells. The choice of the time step size for assured convergence will depend on the wave speed in the dielectric and the minimum grid size.

32.3.2.2 Integral Equation Methods

This method is again based on Maxwell's equation but in the IE format. The IE is further casted in a discretized form appropriate for a numerical solution for the problem. A well accepted method for discretization of the equation is the method of moments (MoM). The MoM procedure is also known as

a general method for numerical solution for both differential and IEs [44]. Here, we refer to the integral process only. This method only requires consideration of the conducting part of the structure and not the entire problem space as in the differential method, so the resultant matrix size is smaller. However, due to non-vanishing coupling of the currents and charges in the problem, the matrix is full leading to longer solution times.

A significant attribute of the integral method is that the equation can be cast in a format that can be interpreted as resistive, capacitive, and inductive elements of the problem. These can then be taken further to construct a complete distributed circuit description of the discrete electromagnetic problem. The resulting circuits are known as partial element equivalent circuits or PEEC [48]. The circuits can then be used in a simulator program [49,50], such as SPICE to analyze the package for its response to electrical stimuli and load conditions.

Depending on how the set of Maxwell's equation is utilized in developing the PEEC model and for the accuracy level supported by the discretized structure, the resultant model can be considered as a full wave solution. This implies that the PEEC model can be used to predict all necessary electromagnetic interactions relevant to the problem of interest.

In a simpler but important application of the IE method using MoM solution techniques, a 2D representation of the signal traces (cross section) can also be efficiently solved. This application is known as a "boundary element method." Although it is a 2D solution, it can provide important information about the package structure such as the per-unit-length capacitance and loop inductance values in the high frequency limit and the characteristic impedance at the point of the 2D section [46]. The solution can be very fast in comparison to the full 3D approach. This technique is offered in many commercial and university codes available today.

32.4 Hygro-Thermo-Mechanical Behavior of Packages

In this section, we discuss the hygro-thermo-mechanical behavior of plastic area-array packages, which have become an enabling technology for future packaging development.

We present an integrated simulation technique to determine stresses in an electronic package arising from moisture diffusion and expansion due to thermal loading. This will enable better predictions of the reliability of the electronic packages and will allow adjustment of package configurations or properties to increase the reliability of the package.

We review results from recent studies using moiré interferometry and show that a considerable advance has been made in understanding the thermo-mechanical behavior of this type of package. In particular, we discuss the role of the underfill and how it reduces the shear strains of the solder balls, but shifts the reliability concern to delamination of the underfill interfaces. An experimental methodology has been described integrating material characterization, experimental measurements, and modeling verification in an attempt to develop a capability to predict the thermo-mechanical behavior of the package. At this time, there seems to be a good agreement in the displacements obtained by experiments and modeling but not in the shear strains. Possible sources of errors are identified and further studies are suggested to improve the moiré measurements and the finite element analysis (FEA).

Electronic packaging serves as an important link between the silicon and the electronic system. It has been developed in parallel with the semiconductors to provide the functional capabilities required to utilize the density and performance built into the integrated circuits. The growth of the area-array package of the last 10 years has proved the capability of this package to support the I/O pad counts and power distribution required by the increase of the device density on the chip. The area-array configuration also improves the electrical performance by reducing the interconnect paths and the signal noise. The development of area-array packages has incorporated new sets of materials, processes, and interconnect structures, which because of specific functional requirements, are not always compatible. This has given rise to manufacturing and reliability problems, which have to be solved to ensure the development of this technology. For this purpose, it is important to have a good

understanding of the materials, processing, and structures and their interplay in controlling the thermal, mechanical, and electrical performance of the package. This section reviews our current understanding on the thermo-mechanical behavior of plastic area-array packages.

The area-array package or the flip-chip solder interconnect connects the active device side of the semiconductor face-down via solder balls on a multi-layered substrate. This technology was first used in the late '60s [55] and continued to develop until the '80s, when it evolved to become high I/O, area-array interconnects for very large-scale integration (VLSI) applications [56]. The development of this technology has been reviewed by Koopman et al., tracing its early development to VLSI applications in the 1980s [57] and in a recent review on the materials and mechanics issues for ultra large-scale integration applications by Wu et al. [16]. The concern for structural integrity due to thermal deformation of the solder balls has been recognized from the beginning of the development of flip-chip packages. The problem arises because of the mismatch of the CTE between the chip and the substrate, which are 3 ppm/°C for Si and about 10 ppm/°C for a ceramic substrate and higher for a plastic substrate. During thermal processing or in circuit operation, thermal cycling induces shear strains on the solder balls. This strain increases with the distance from the neutral plane, reaching a maximum at the outermost solder row on the chip. The problem this generates is solder fatigue in the highly strained balls and eventual interconnect failure. This problem can be solved in two ways. The first is to choose materials in the substrate board to match the thermal expansion of the silicon, thus eliminating the driving force for shear deformation. The second method is to use a compliant substrate so that it, instead of the solder ball, will deform to accommodate the thermal strain. Indeed, the first flip-chip package was made with a thin and flexible polyimide substrate in a "decal" configuration [58]. These solutions, however, are inadequate in meeting the demands for large chip size and low cost packages. Most of the low-cost plastic substrates have a large CTE compared with Si—being typically 20–30 ppm/°C. Even if this CTE could be reduced by 50%, its difference with Si can still generate a large shear strain in the solder balls. During the early development of flip-chip packages, low-cycle thermal fatigue in solder balls dominated the reliability issue.

A major breakthrough came in the early '90s with the introduction of the underfill encapsulant, which is dispensed by a capillary action to fill the gap between the chip and the substrate [59]. The encapsulant serves as a compliant buffer reducing the shear strain of the solder balls by coupling the thermal mismatch into bending of the substrate, resulting in a significant improvement of the fatigue life of the solder balls. This innovation provided a paradigm shift in packaging development, making it possible to design plastic flip-chip packages. These packages could be designed to meet the requirements for high I/O counts and large chip size for deep-submicron technology. Although the use of the underfill encapsulant eliminates to a large extent the problem of low-cycle solder fatigue, a number of reliability issues remain. First, of all, there are interfacial cracks induced by material reactions during solder reflow between the solder and the top and the bottom metallization layers [60]. Second, the underfill layer shifts the local shear strain from the solder balls to the underfill/die and the underfill/substrate interfaces. This strain can cause delamination at these interfaces to become a major concern for the structural integrity of the package. And third, the requirements for the underfill properties become more stringent due to increasing I/O density and decreasing gap height. These requirements lead to a demand for underfills with good flow as well as thermo-mechanical matching to the die and the substrate.

While addressing these issues has stimulated considerable recent effort in materials and processing studies, the discussion in this section is concentrated on the thermo-mechanical behavior of the area-array package. This discussion is based primarily on results from recent moiré interferometry studies aiming to develop a methodology to integrate material characterization, experimental measurement, and computer modeling for evaluation of plastic area-array packages. We will show how these studies can provide useful feedback for materials and processing optimization for packaging development. This article is organized into three sections. First, the thermal deformation of the flip-chip package is described. This is followed by discussions on measurements of the underfill properties and then the modeling verification of the measured thermal deformation.

32.4.1 Solder Joint Fatigue Failure

The degree of complexity of an electronic package is identified by three distinct assembly levels: (1) attachment of the chip (die) to the chip carrier (substrate), (2) attachment of the chip carrier and the die to the PCB, and (3) the attachment of the PCB to the other PCBs through edge connects or to other devices through cables. The SMT, commonly used in the electronics industry, enables the attachment of the chip directly to the PCB. The flip-chip technology, an advanced form of SMT, for attaching an upside down chip equipped with solder bumps on its surface to the substrate and then to the PCB with BGA connection is described in Figure 32.3. In order to enhance its mechanical performance, a numerical simulation of the mechanical behavior of an electronic package is essential. However, the fidelity of the simulation is dependent upon accurate numerical modeling, the material models, and the failure criterion. Failure mechanisms in an electronic package can be classified as intrinsic and extrinsic. The intrinsic failures arise from the process-related defects in silicon wafer substrates, dielectrics, insulating films, interconnecting films between devices, and components. The extrinsic failures arising from the external mechanical and thermal loads, and environmental conditions, may include chip fracture, solder joint fracture, moisture-induced swelling and cracking of the encapsulation, corrosion, creep, and fatigue of solder joints.

The most commonly accepted models employing the FEM for thermal fatigue reliability analyses can be grouped as: (1) nonlinear slice model, (2) global model with linear super elements and non-linear solder, (3) linear global model with nonlinear submodel, (4) nonlinear global model with a nonlinear submodel and, (5) nonlinear global model. In the global analyses with a submodel, the displacement fields obtained from the analyses are extrapolated for different temperatures and applied as boundary conditions in the submodel of the solder joint with a refined mesh. In the nonlinear analyses, four thermal cycles are sufficient to achieve a steady-state response (stable hysteresis loop). The details of the all of these modeling approaches are given by Madenci et al. (2002).

The deformation behavior of solder is dependent on temperature and time. Therefore, the deformation of solder is described by time-dependent constitutive laws of viscoplasticity or time-dependent creep combined with time-independent plasticity. In viscoplastic deformation, the elastic region is bounded by a so-called static yield surface in stress space and all inelastic deformations are time-dependent. Also, the inelastic deformations occur at all non-zero stress values. The viscoplastic constitutive law does not distinguish the plastic strains from those of creep. Unlike the time-independent plasticity law, the viscoplastic constitutive law does not rely on an explicit yield surface, and the loading and unloading criterion. Instead, it utilizes an internal state variable, representing the resistance of the material to inelastic deformations. Creep deformation is time and temperature dependent and the time-independent plastic deformation results in plastic strains depending on the yield surface, and loading and unloading criterion.

There exist many thermal fatigue life prediction models for determining the solder joint reliability of electronic packages, each with its own merits. These models are dependent on the package-level, statistical (empirical) failure parameters associated with a key parameter of the structure, such as viscoplastic strain energy or matrix creep strain calculated using FEA. The leading failure indicators for the correlation of thermal fatigue life are based on strain ranges such as the total, inelastic, and matrix creep strains (see Iannuzzelli et al. and Syed) [82,83] and the viscoplastic strain energy density increment by Darveaux [84]. One of the widely accepted failure criteria introduced by Darveaux [85] for thermal fatigue-life correlation is based on a relationship in terms of the volume-weighted-average inelastic work density increment. A complete description and application of this criterion to both the thermo-mechanical and mechanical bending fatigue life prediction is given by Madenci [87].

32.4.2 Hygro-Thermo-Mechanical Analysis

The polymeric materials within integrated circuit packages absorb moisture when they are exposed to humid air or other forms of moisture. Electronic packages come in contact with humid air, while being

packaged, transported and stored. Polymer materials expand when they absorb moisture. The polymers continue to absorb moisture and expand as long as they are in contact with humid air, until they are fully saturated. The amount of moisture that a polymer can hold per unit volume is dependent upon temperature and the amount of moisture in the air to which it is exposed.

If a polymer is attached to a non-polymeric material, or a polymeric material with different properties, the difference in expansion between the materials causes stresses in the package. These stresses, referred to as hygroscopic stresses, increase the effect of thermal stresses, when the package is subjected to high temperature solder reflow, where the temperature of the package increases approximately 190°C in a matter of seconds.

In addition to the thermal and hygroscopic stresses within the package, a third type of stress induced by vapor pressure within the polymer must be considered (Figure 32.4). As the moisture disseminates into the polymer material, it is trapped in microvoids within the material. As the temperature of the package rises during reflow, the vapor pressure within the voids increases the stresses within the package leading to delamination and cracking (Wong 1998).

It has been shown by Kitano [86] and Wong et al. [89] that delamination or cracking at the interface between the die and the die-attach layer is one of the primary failure mechanisms in plastic integrated circuit packages. This effect can also weaken the package to the point where other mechanical failures can occur. Another likely scenario is that the voids in the material may lead to corrosion within the electrical conduits, eventually causing electrical failures.

In summary, there are a multitude of variables working against package reliability. Thermal-mechanical and hygro-mechanical stresses as well as stress caused by the vapor pressure within micro-voids in the material; all combine to increase the chances of delamination or cracking occurring at critical interfaces. These concerns will be further aggravated by the regulations currently being implemented that require the use of lead free solders in the electronic packaging industry. Lead free solders melt at a higher temperature so the temperature of the package will rapidly increase from room temperature to approximately 210°C. The 20°C increase in reflow temperature will significantly affect the maximum stresses within the package.

Techniques for in-situ measurement of moisture concentration and stresses within electronic packages do not exist due to the small length scales. Consequently, numerical simulation techniques have become an essential part of electronic packaging design and manufacturing in order to ensure acceptable reliability. This will enable better predictions of the reliability of the electronic packages and will allow adjustment of package configurations or properties to increase the reliability of the package. In order to capture accurate mechanical behavior through numerical analysis, accurate material properties of the constituent materials must be available.

The moisture diffusion modeling is performed using the thermal-moisture analogy (Wong [88]). In this thermal-moisture analogy, the transient heat diffusion equation is used to simulate the diffusion of

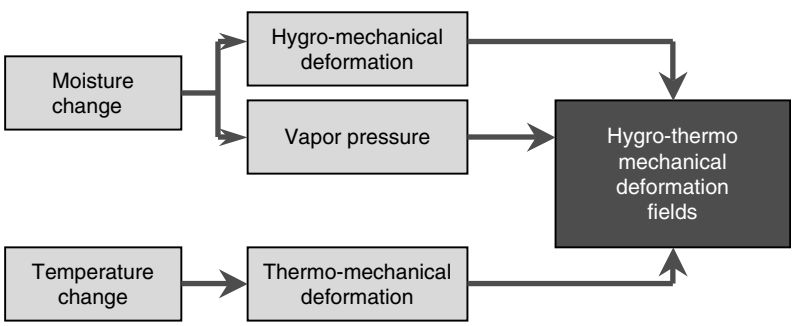


FIGURE 32.4 Interactive hygro-mechanical and thermo-mechanical simulation flow.

moisture within the material. The governing equation for heat and moisture diffusion are given by:

$$\frac{\partial T}{\partial t} = \alpha \nabla^2 T$$
$$\frac{\partial C}{\partial t} = D \nabla^2 C$$

in which T and C represent the temperature and moisture concentration, while the parameters α and D denote the thermal and moisture diffusivity, respectively.

Although these equation are of the same type, this analogy cannot be used directly due to the lack of moisture concentration continuity along material interfaces. In order to ensure that continuity is enforced along the material interfaces, a new parameter is introduced called the "wetness" parameter. The moisture concentration C is normalized with respect to the saturated moisture concentration C_{sat} leading to a dimensionless wetness variable in the form:

$$w = \frac{C}{C_{\text{sat}}}$$

When the wetness (w) is equal to zero (0), the material is completely dry and when w is equal to one (1), the material is fully saturated. Substitution of w into the moisture diffusion equation results in the governing equation for wetness:

$$\frac{\partial w}{\partial t} = D \nabla^2 w$$

The moisture diffusion is analogous to that of thermal diffusion when the wetness normalization is employed. Therefore, transient heat diffusion FEA can be employed to perform moisture diffusion simulations. In the FEA, w is analogous to temperature T when the substitutions shown in Table 32.4 are made.

The weight of water in each element can be computed by averaging the moisture concentration in the form

$$\frac{\text{Weight of water}}{\text{Element}} = \left(\frac{1}{N} \sum_{i=1}^{N} w_i \right) C_{\text{sat}} V$$

in which N is the number of nodes per element and V represents the volume of the element. The wetness at the ith node is indicated by w_i. The weights of the elements are combined to determine the total weight of the composite structure.

In order to perform the hygro-mechanical analysis described in detail by Madenci et al. [87], first the moisture diffusion analysis is conducted leading to the determination of the moisture concentration distribution in the specimen during the desorption process (see Figure 32.5). Then, the deformation analysis is conducted based on the transient moisture concentration and temperature distributions.

TABLE 32.4 Correspondence Table for Thermal/Moisture Concentration

Property	Thermal	Moisture
Primary variable	Temperature, T	Wetness, w
Density	ρ (kg/m^3)	1
Conductivity	κ (W/m °C)	$D \cdot C_{\text{sat}}$(kg/s m)
Specific capacity	c (J/kg °C)	C_{sat}(kg/m^3)

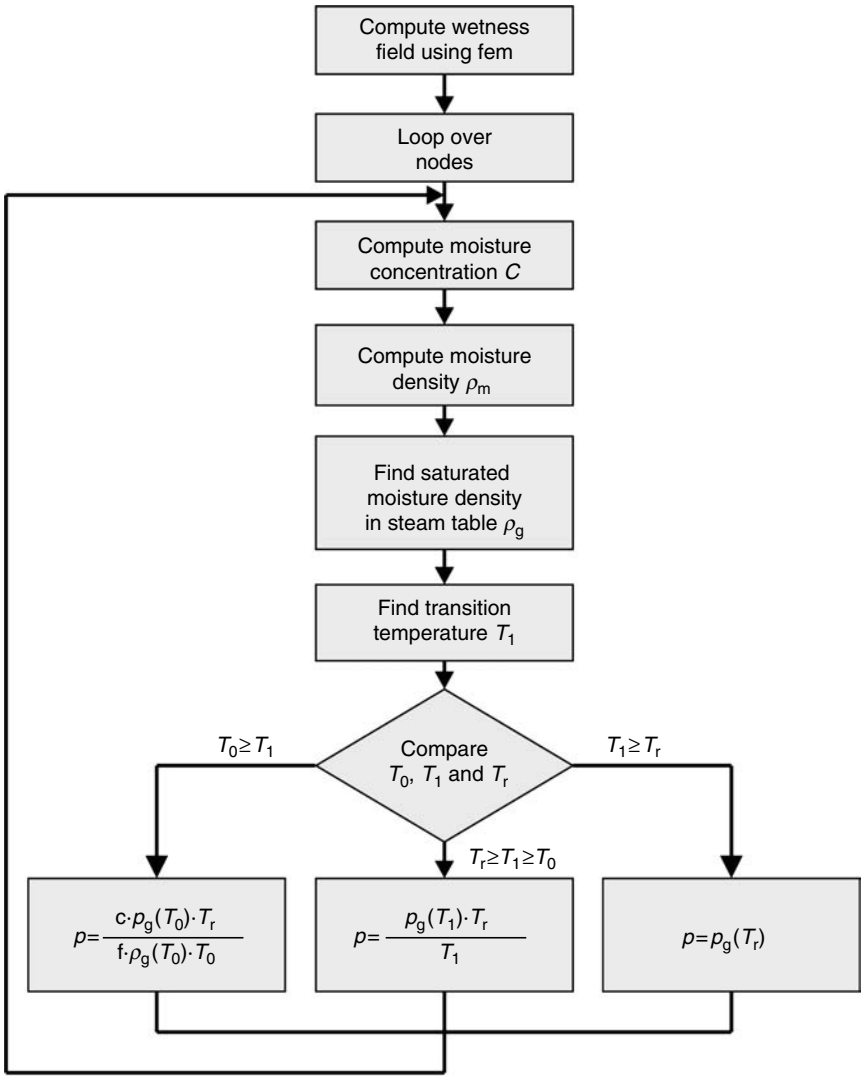

Compute wetness field using fem
Loop over nodes
Compute moisture concentration C
Compute moisture density ρ_m
Find saturated moisture density in steam table ρ_g
Find transition temperature T_1
Compare T_0, T_1 and T_r
$T_0 \geq T_1$
$T_1 \geq T_r$
$T_r \geq T_1 \geq T_0$
$p = \frac{c \cdot p_g(T_0) \cdot T_r}{f \cdot \rho_g(T_0) \cdot T_0}$
$p = \frac{p_g(T_1) \cdot T_r}{T_1}$
$p = p_g(T_r)$

FIGURE 32.5 Hygro-mechanical analysis flow.

Based on the moisture concentration distribution, the vapor pressure is computed as shown in the following flow of computations, Wong et al. [90].

32.4.3 Thermal Deformation of Flip-Chip Packages

An environmental moiré interferometry technique was used for in-situ mapping of the thermal deformation in flip-chip packages [61]. Moiré interferometry is a whole-field optical technique with high sensitivity and spatial resolution for measuring in-plane displacements and strains [62,63]. An environmental chamber was used together with a Portable Engineering Moiré Interferometer developed by IBM to acquire in-situ deformation patterns as a function of temperature in a controlled atmosphere [64]. The underfilled flip-chip-on-board package used in this study had double row peripheral solder bumps with a total 340 I/O. The chip had dimensions of 12.9 mm × 12.9 mm × 0.62 mm and was mounted on a 1.5 mm thick substrate of a FR4 PCB embedded with copper lines. The gap between the

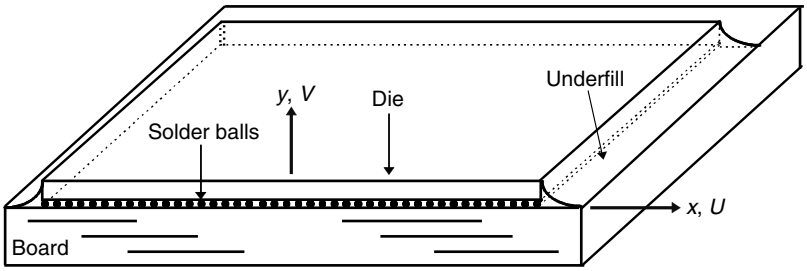


FIGURE 32.6 Schematic drawing of the underfilled flip-chip-on-board package under study.

chip and the substrate was 75 μm and the solder bumps, made of eutectic solder, had a 250 μm pitch. The underfill materials were silica-filled liquid epoxy.

The actual specimen, shown schematically in Figure 32.6 was cross-sectioned (see Figure 32.7) for moiré studies with a grating attached at 102°C. This grating served as the starting temperature for tuning the null moiré displacement fields. As the temperature decreased from 102 to 22°C, fringe patterns for the horizontal displacement (U) field and the vertical displacement (V) field evolved as shown in Figure 32.8 and Figure 32.9, respectively. The interference fringes represented constant displacement contours with each fringe separated consecutively by a displacement of 0.417 μm. With increasing cooling, the increase in fringe density showed that all the components contracted more and more in both the horizontal and the vertical directions. This resulted in increasing downward bending of the whole package, as revealed by denser inclined fringes in the V displacement fields. The low fringe density in the silicon chip showed a relatively small deformation, which was due to the low CTE and high Young's modulus of the material. In the PCB, the vertical contractions were much greater than the horizontal contractions, as shown by the higher fringe gradient in the V-field fringes, which was caused by the anisotropic CTE properties of the board. Due to its composite layer structure, the circuit board exhibited irregular fringes. However, by tracing the fringes carefully, one can determine the number of composite layers that make up the board structure.

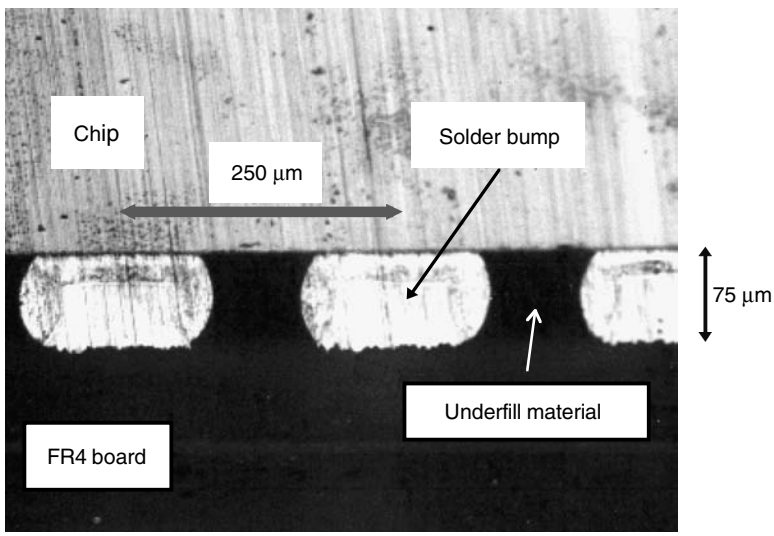


FIGURE 32.7 Cross-sectional picture of the underfilled flip-chip interconnections.

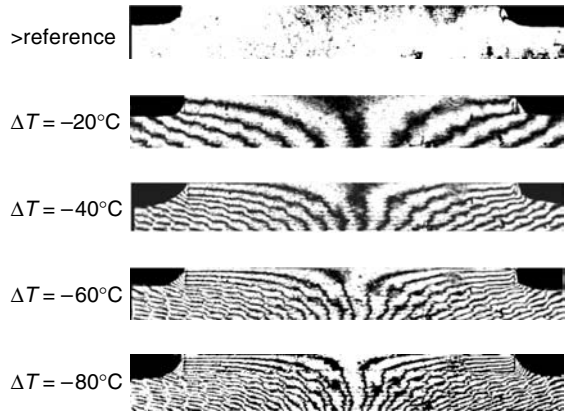


FIGURE 32.8 Horizontal (U) displacement fields and under thermal loading conditions (reference temperature $T = 102°C$).

In the horizontal direction, the CTE mismatch resulted in a relative displacement between the chip and the board, which increased with the distance from the geometric center of the package. To provide a better view for the solder/underfill area between the chip and the board, the U and V fields at 22°C were captured with a greater magnification. These results are shown in Figure 32.10a and b, where the increase in fringe densities in the solder/underfill layer with distance from the center can be readily observed. At the ends of the die, the relative displacement reached its maximum value, indicating the highest stress and strain in the outermost solder balls.

The overall deformation can be quantitatively determined by analyzing the spatial distribution of the fringes. With a displacement sensitivity of 0.417 μm per fringe order, the solder balls, which are approximately 75 μm × 150 μm in cross section, carry very few fringes even at high strain. To provide more detailed information of the strain distributions, carrier fringes [65] were employed to increase the data points for fringe analysis to determine the average shear strain across the solder balls. This is demonstrated for the three rightmost solder ball in Figure 32.11a for the U fields with carrier fringes together with the V fields in Figure 32.11b. The V fringes show a significant deformation in the

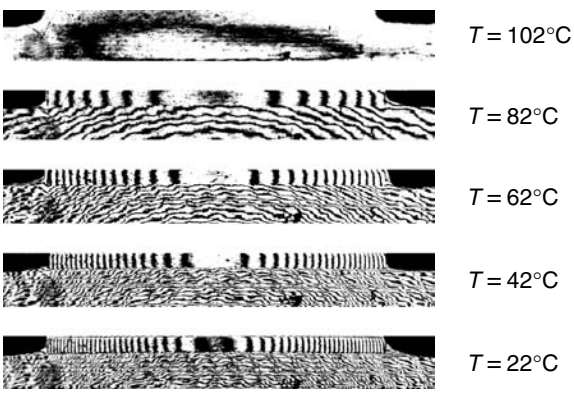


FIGURE 32.9 Vertical (V) displacement fields under thermal loading conditions (reference temperature $T = 102°C$).

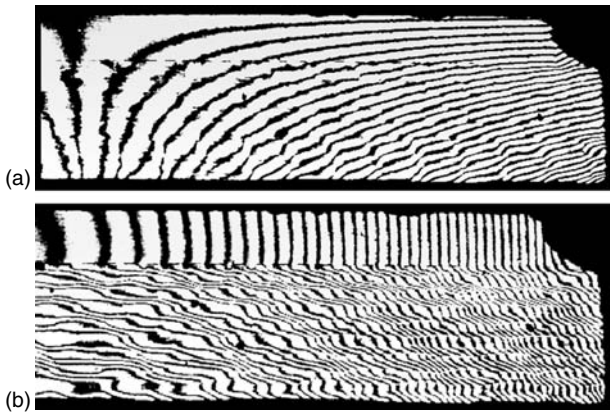


FIGURE 32.10 (a) U and (b) V fringe patterns under thermal loading of $\Delta T = -80°C$ over the right half of the package.

underfill/solder layer caused by the downward bending of the package balls under a thermal loading of $\Delta T = -80°C$. In comparison, the U field carrier fringes are evenly distributed, indicating relatively uniform horizontal displacements for these solder balls due to the presence of the underfill.

The components of the shear strains of the solder balls, $\partial U/\partial y$ and $\partial V/\partial x$, were calculated using the relative horizontal and vertical displacements divided by the height and width of the solder balls, respectively. The results are plotted in Figure 32.12 for the rightmost solder ball as a function of temperature. A linear behavior of the shear strains was observed. It is clear that the $\partial U/\partial y$ component coming from thermal mismatch between the die and the board is substantially canceled by $\partial V/\partial x$, caused by the board bending, resulting in a much smaller total shear strain. Thus, the underfill plays an important role in reducing the shear strain of the solder balls by inducing bending of the structure to cancel the thermal mismatch between the die and the substrate. The substrate bending, however, shifts the shear strain to the die/underfill and the underfill/substrate interfaces and generates additionally peeling stresses at these interfaces. This interfacial stress can cause interfacial delamination instead of solder fatigue to become a primary reliability concern. The characteristics of the interfacial shear strains will be discussed later in modeling verification.

This effect was further studied using two underfills, which have the same resin chemistry but different filler contents. Except for the underfills, identical test packages were used. Results from moiré measurements found that these underfilled packages have essentially the same $\partial U/\partial y$ over solder/underfill area, which is to be expected, since the horizontal deformation comes from thermal mismatch between the die and the substrate and should be independent of the underfill. However, the underfilled package with

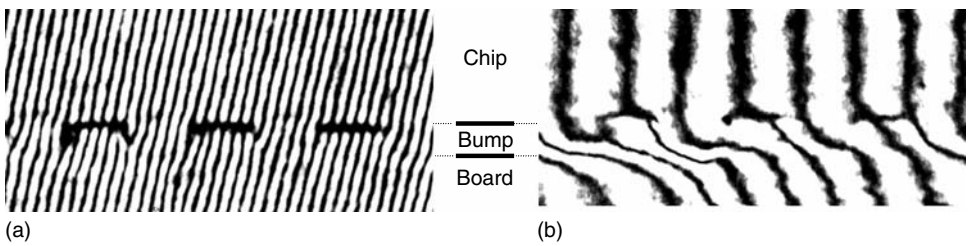


FIGURE 32.11 (a) U displacement field with carrier fringes of contraction; (b) V displacement field over the three rightmost solder balls.

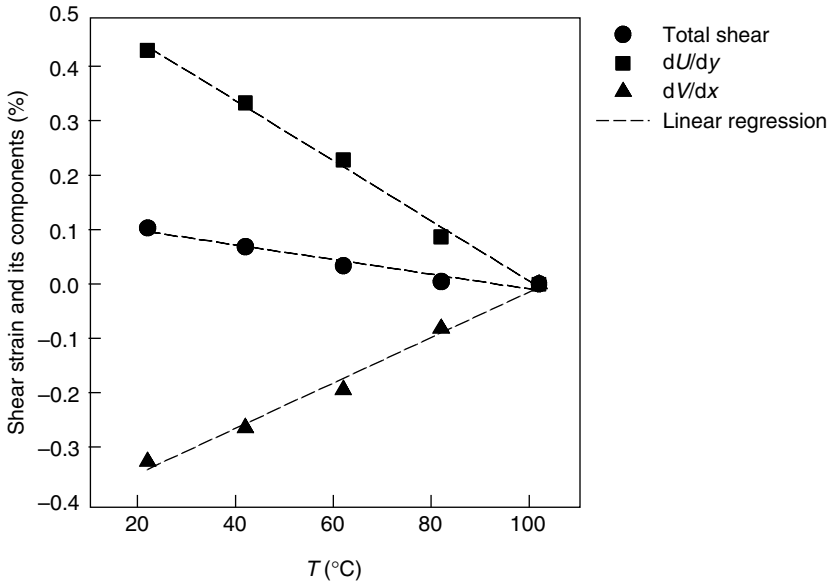


FIGURE 32.12 Total shear strain and its components across the rightmost solder ball as a function of temperature.

the higher filler content exhibited a higher $|\partial V/\partial x|$ value. The larger deflection of the package is due to the higher filler content, 70% vs. 55%, which reduces the CTE but increases the elastic modulus of the underfill. (See next section for characterization of the underfill properties.) As a result, the total shear strain over the solder/underfill area is reduced. Results from this study (Figure 32.13) suggest that the deformation behavior of the underfilled package can be changed by varying the underfill properties, leading to an approach to optimize materials for area-array packages.

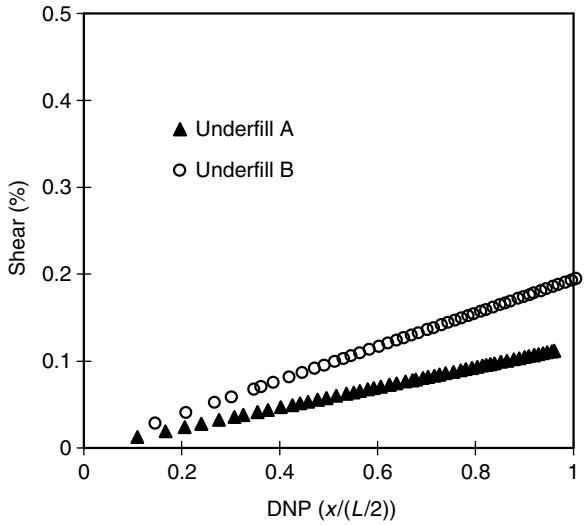


FIGURE 32.13 Total shear strain across solder/underfill area as a function of normalized distance to the neutral point (DNP) under thermal loading of $\Delta T = -80°C$.

32.4.4 Characterization of Material Properties

Material properties are required for computer modeling of the hygro-thermo-mechanical behavior of packages. Of particular interest are moisture diffusivity, solubility, coefficient of moisture expansion (CME) and CTE and, elastic modulus of polymeric materials used in electronic packages. The CTE and elastic modulus, which can be determined using moiré interferometry in combination with a bending beam technique [66]. The bending beam technique was applied to a composite structure consisting of a silicon substrate with a shallow trough containing a layer of the underfill. By measuring the bending of the composite during curing and thermal cycling, the stress generated in the underfill can be determined. The temperature dependence of the stress can provide a direct measure of the curing behavior and the glass transition temperature. Under thermal cycling, the slope of the stress vs. temperature is directly proportional to the product of the biaxial modulus and the CTE difference between the underfill and silicon. When it is used in combination with moiré interferometry, which measures CTE directly, the elastic modulus can be deduced.

This is illustrated in Figure 32.14a and b of the U and V displacement fields of the PCB used in the flip-chip package. The zigzag nature of fringes in the PCB reveals the heterogeneous deformation of the multi-ply glass/epoxy composite. Since, the PCB has a much larger CTE in the out-of-plane direction than the in-plane direction, the V field has a higher fringe density than the U field. The CTEs determined in the temperature range of 102°C–22°C are 14.7 ± 0.7 and 51.2 ± 3.0 ppm/°C for the in-plane and the out-of-plane directions, respectively. The material properties of the two underfills used in our study have been determined by combined moiré and bending beam measurements, and the results are summarized in Table 32.5. The results demonstrated that the underfill properties can be modified by changing the filler content, which in turn affects the thermal deformation of the flip-chip packages.

The measurement of moisture diffusivity, solubility and, CME of polymeric materials are dictated by JEDEC standards JESD22-A120 and JESD72. The first standard provides test methods for the measurement of moisture diffusivity and water solubility in polymer materials used in integrated circuits, while the second provides test methods for the evaluation of polymeric materials and includes a discussion of CTE. The CMEs of the packaging materials are determined by modifying the thermal mechanical analysis (TMA) and thermal gravitational analysis methods. These methods provide the change in length and mass, respectively, of specimens as moisture is desorbed from a preconditioned (fully saturated) sample. Method to determine the CME of the materials as moisture is simultaneously absorbed requires two nearly identical specimens. This method is also a modification of the TMA test performed inside an environmental chamber. The details of the measurements and test setup are described by Madenci et al. [87].

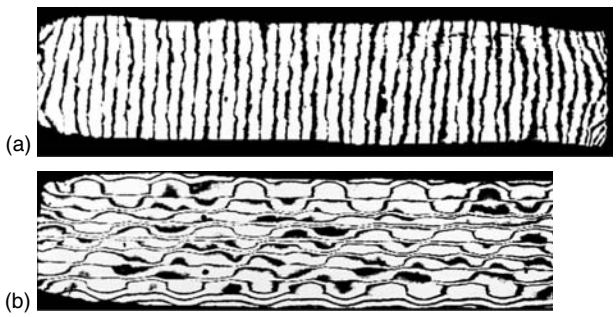


FIGURE 32.14 Moiré fringes: (a) U field and (b) V field for a printed circuit board specimen under a thermal loading of $\Delta T = -80$°C.

TABLE 32.5 Underfill Materials Property Data

Underfill	Filler Content (%)	Coefficient of Thermal Expansion (CTE) (ppm/°C)	Modulus (GPa)
Underfill A	67–72	22	5
Underfill B	55	35	3

32.4.5 Modeling Verification for Flip-Chip Packages

So far the discussion has been concentrated on results from experimental studies based on moiré interferometry to investigate the underfill materials and the thermal deformation of plastic area-array packages. This was supplemented by modeling verification in an attempt to develop a capability to predict the deformation behavior for plastic area-array packages. A 3D finite element model was applied to analyze the deformation, strain, and stress behavior of the test package. The geometry of the model was based on the test package used for moiré interferometry study. The model, shown in Figure 32.15 uses a plane symmetry resulting in one half of the moiré specimen being simulated. All materials were assumed to behave linear elastically with temperature independent material properties. Perfect adhesion was assumed to exist for all material interfaces. The silicon die and the underfill were assumed to be isotropic and the FR4 board was assumed to be orthotropic. The solder balls, which are located peripherally did not contribute much to the deformation behavior of the package compared to the contribution of the underfill; so they were not modeled separately [67].

For modeling verification, we concentrated on the displacements and strains at the die/underfill and underfill/substrate interfaces, as these show the highest strain concentration and are most prone to delamination. The fitting of the moiré U and V displacements was based on function forms from Suhir's shear stress solution, which was deduced using linear elastic beam theory [68]. The results show a good overall agreement for the U and V displacements measured by moiré and predicted by the FEA at the die/underfill and the underfill/substrate interfaces (Figure 32.16). A similar agreement was obtained for the displacements at the underfill/substrate interface. This result suggests that the functional forms used to fit the moiré displacements and the representation of the joint region as a homogeneous layer are reasonable assumptions.

The agreement between the deformation strains components, $\partial U/\partial y$ and $\partial V/\partial x$, deduced from the moiré and FEA results are also satisfactory, as shown in Figure 32.17a and b. for the die/underfill interface as a function of the distance to the neutral point (DNP). For both $\partial U/\partial y$ and $\partial V/\partial x$, we see an agreement in the rising trend with increasing DNP in the moiré and FEA results, although there is some deviation, which reaches maximum under the die edge in both the cases. The shear strains in the joint area obtained from moiré and FEA are summarized in Figure 32.18. The shear strains at the die/Underfill (UF) interface calculated from moiré displacements are represented by solid triangles, while

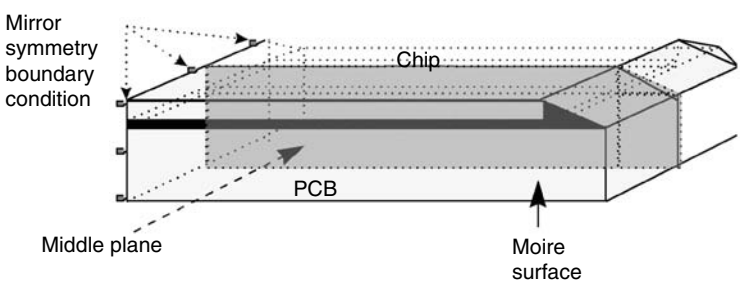


FIGURE 32.15 The 3D finite element model.

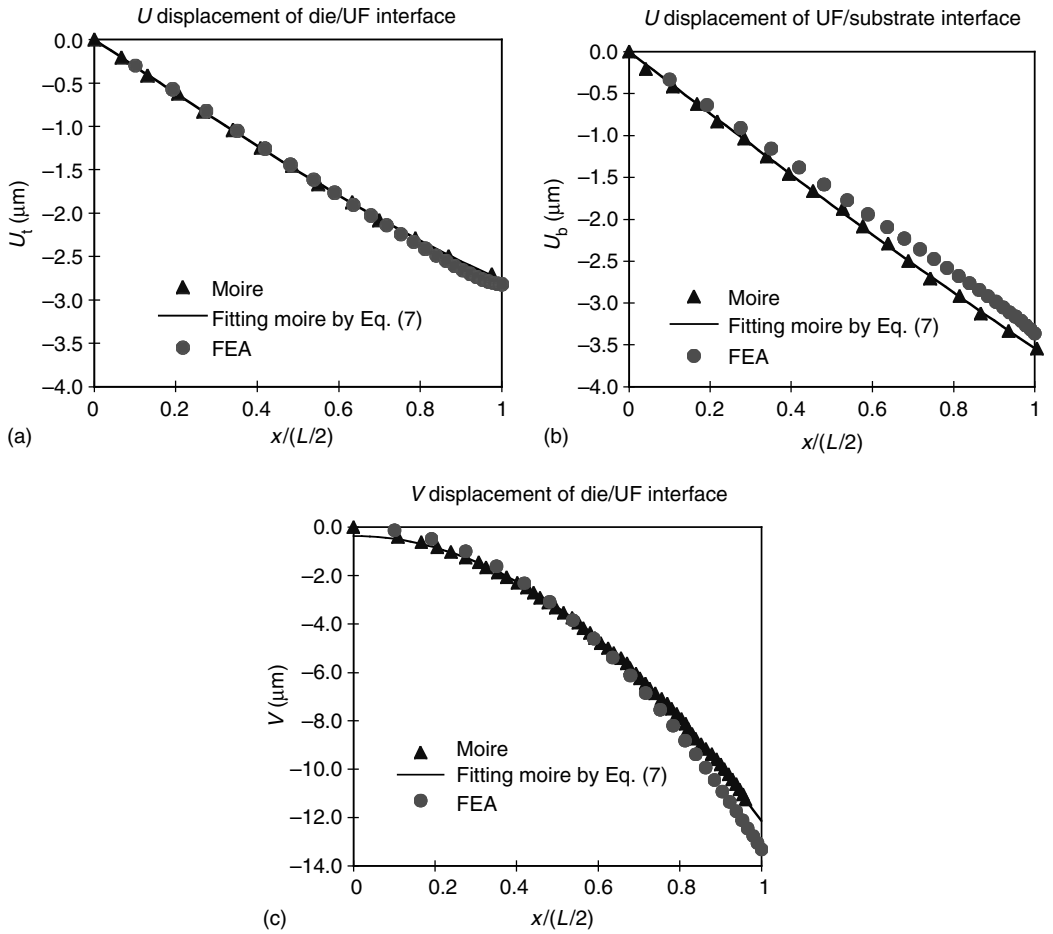


FIGURE 32.16 Moiré and finite element analysis (FEA) in comparison: (a) U displacement of die/UF interface; (b) U displacement of UF/substrate interface; and (c) V displacement of die/UF interface.

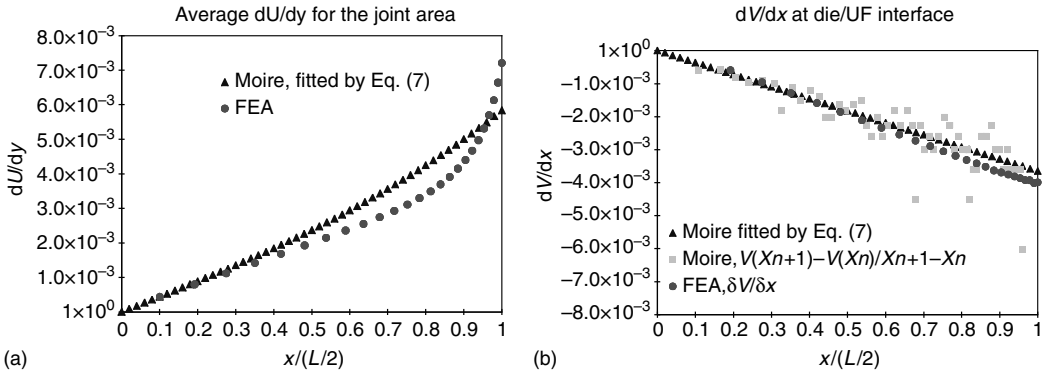


FIGURE 32.17 Comparison of shear strain components of (a) $\partial U/\partial y$ and (b) $\partial V/\partial x$ obtained by moiré and FEA.

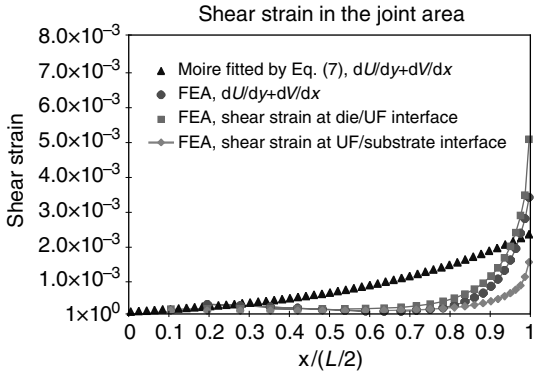


FIGURE 32.18 Comparison of total shear strains in the joint area obtained by moiré and FEA.

those from FEA are represented by solid circles. Additionally, the shear strain obtained by FEA is shown for the die/UF interface (solid squares) and the UF/substrate interface (solid diamonds). One can see that the shear strain at the die/underfill interface is more than two times of the shear strain at the underfill/substrate interface. Thus, of these two interfaces, the die/underfill interface is more prone to delamination under thermal loading.

The overall agreement between the moiré and FEA results for the shear strains at the die/underfill interface is not satisfactory, although both show a consistent rising trend as one moves outward from the center to the edge of the die. Possible sources responsible for the discrepancy can be attributed to the experimental uncertainities and modeling assumptions. There are experimental uncertainties in obtaining and fitting the moiré data to deduce the deformation strains. We believe that a higher sensitivity would help to increase the number of data points that can be used and thus eliminate some of the uncertainties. The modeling assumptions, such as perfect adhesion, linear elastic behavior, and temperature-independent material properties can also contribute.

32.4.6 Summary of Thermo-Mechanical Behavior

In this section, we have discussed the thermo-mechanical behavior of plastic area-array packages, which has become an enabling technology for future packaging development. The change to area-array interconnect with high I/O counts and power dissipation has made thermal deformation an important concern for package reliability. We have reviewed results from recent studies using moiré interferometry and shown that a considerable advance has been made in understanding the thermo-mechanical behavior of this type of packages. In particular, we discussed the role of the underfill and how it reduces the shear strains of the solder balls, but shifts the reliability concern to delamination of the underfill interfaces. An experimental methodology has been described integrating material characterization, experimental measurements, and modeling verification in an attempt to develop a capability to predict the thermo-mechanical behavior of the package. At this time, there seems to be a good agreement in the displacements obtained by experiments and modeling, but not in the shear strains. Possible sources of errors have been identified, and further effort is required to improve the moiré measurements and the FEA. There are many excellent references for further review of thermo-mechanical topics of particular interest are other studies of moiré interferometry [69–71], stress measurement and modeling verification [72], interfacial fracture on area-array packages [73–78], and analytical stress models for multi-layered microelectronics structures [59,79–81].

Looking ahead, the technology will be driven by increasing chip size, interconnect density, and power, which will require development of area-array packages with a small gap, more I/O, and greater power

dissipation. These changes together with demands for cost reduction and fast product cycle will present serious challenges to the industry. Better experimental methodology and computer modeling capability will be needed for future investigation of plastic area-array packages.

References

1. *The International Technology Roadmap for Semiconductors*. 2005 ed., (http://www.itrs.net).
2. Tummala, R. R., E. J. Rymaszewski, and A. G. Klopfenstein, eds. *Microelectronics Packaging Handbook Part I—Technology Drivers*. New York: Chapman & Hall, 1997.
3. Seraphim, D. P., R. C. Lasky, and C.-Y. Li. *Principles of Electronic Packaging*. New York: McGraw-Hill, 1989.
4. Bohr, M. T. "Interconnect Scaling—The Real Limiter to High Performance ULSI." In *Proceedings of the 1995 IEEE International Electronic Devices Conference*, 241–2, 1995.
5. Harman, G. *Wire Bonding in Microelectronics Materials, Processes, and Reliability, Yield*. 2nd ed. New York: McGraw-Hill, 1997.
6. Chang, C. S. "Heuristic Equations for Semiconductor and Packaging Technology Evolution Projection." *IBM Unclassified Technical Report, TR 01.C772*, November 21, 1994.
7. Chang, C. S., and A. Oscilowski. "U.S. Packaging Trends." *SEMICON/Japan '96, Proceedings, International Packaging Strategy Symposium*, 1-1 to 1-17. Tokyo, Japan, December 3, 1996.
8. Totta, P. A., S. Khadpe, N. G. Koopman, T. C. Reiley, and M. J. Sheaffer. "Chip-to-Package Interconnections." In *Microelectronics Packaging Handbook*, edited by R. R. Tummala, E. J. Rymaszewski, and A. G. Klopfenstein *Part II—Semiconductor Packaging*, II-129–II-283. New York: Chapman & Hall, 1997.
9. Yamamoto, H., A. Fujisaki, and S. Kikuchi. "MCM and Bare Chip Technology for a Wide Range of Computers." In *Proceedings of 46th Components and Technology Conferences*, 133–8. Orlando, FL, May 28–31, 1996.
10. Young, B. "Figures of Merit for Package Electrical Roadmaps." *IEEE Trans. Comp., Packag. Manuf. Technol.* 21 (1998): 281–5.
11. Choi, K. L., N. Na, and M. Swaminathan. "Characterization of Embedded Passives Using Macromodels in LTCC Technology." *IEEE Trans. Comp., Packag. Manuf. Technol.* 21 (1998): 258–68.
12. Cangellarias, A., and J. Prince. "Modeling and Simulation for Mixed Signal Package Design." *Adv. Electron. Packag.* 19, no. 1 (1997): 497–507.
13. Huang, C., M. Celik, and J. L. Prince. "Simultaneous Switchuing Noise Simulation for Thin Film Packages Using Macromodeling Technique." In *Proceedings of Electronic Components Technology Conference*, 747–51, 1996.
14. Landman, B. J., and R. L. Russo. "Pin vs. Block Relationships for Partitions of Logic Graphs." *IEEE Trans. Comput.* C20, no. 12 (1971): 1469–79.
15. Semiconductor Industry Association. *The National Technology Roadmap for Semiconductors*. (1994). San Jose, CA: Semiconductor Industry Association, 1994.
16. Wu, T. Y., Y. Tsukada, and W. T. Chen. "Materials and Mechanics Issues in Flip-Chip Organic Packaging." In *Proceedings of 46th Electronic Components and Technology Conferences*, 524–34. Orlando, FL, May 28–31, 1996.
17. Gamota, D., and C. Melton. "Reflowable Material Systems to Integrate the Reflow and Encapsulant Dispensing Process for Flip Chip on Board Assemblies." *Int. Conf. Electron. Assem. Matl. Proc. Chall.* Atlanta, GA, IPC-TP-1098, 1996; Gamota, D., and C. Melton. "Materials to Integrate the Solder Reflow and Underfill Encapsulation Processes for Flip Chip on Board Assembly." *IEEE Trans. Comp. Packag. Manuf. Technol., Part C* 21, no. 1 (1998): 57; Wong, C. P., and D. F. Baldwin. "Novel No Flow Underfills for Low-Cost Flip-Chip Applications: Materials and Processes." *Future Circuits Int.* 3 (1998): 67–70.
18. Zorba, D., and M. Edwards. "Review of Underfill Encapsulant Development and Performance of Flip Chip Applications." *ISHM '95 Proceedings*, 354–8, 1995.

19. Dai, X., and P. Ho. "Thermo-Mechanical Deformation of Underfilled Flip-Chip Packaging." In *Proceedings of 1997 International Electronic Manufacturing Technology Symposium*. Austin, TX, October 13–15, 1997.
20. Goodson, K., K. Kurabayashi, and F. Pease. "Improved Heat Sinking for Laser-Diode Arrays Using Microchannels in CVD Diamond." *IEEE Trans. Comp., Packag., Manuf. Technol., Part B: Adv. Packag.* 20, no. 1 (1997): 104–9.
21. Mizumoto, S., Y. Tsukada, and I. Shishido. "Analysis of Stress Condition for Organic Multi-Chip Package." In *Proceedings of International Intersociety Electronic and Photonic Packaging Conference (INTERpack '97)*, 113–7. Kohala Coast, HI, June 15–19, 1997.
22. Chu, R. C., U. P. Hwang, and R. E. Simons. "Conduction Cooling for an LSI Package: A One Dimensionsl Approach." *IBM J. Res. Dev.* 26, no. 1 (1982): 45–55.
23. Lau, J., C. P. Wong, J. L. Prince, and W. Nakayama. *Electronic Packaging: Design, Materials, Processing and Reliability*, 498. New York: McGraw-Hill Company, 1998.
24. Senthinathan, R., and J. L. Prince. *Simultaneous Switching Noise of CMOS Devices and Systems*, 204. Boston, MA: Kluwer Academic Publishers, 1994.
25. Jiao, C., A. C. Cangellaris, A. Yaghmour, and J. L. Prince. "Sensitivity Analysis of Multiconductor Transmission Lines and Optimization for High-Speed Interconnect Circuit Design." In *Proceedings of 49th IEEE Electronic Components and Technology Conference*, 480–7. San Diego, CA, June 1999.
26. Kielkowski, R. *Inside SPICE*. 2nd ed. New York: McGraw-Hill, Inc., 1998.
27. Taflove, A., and K. R. Umashankar. In *The Finite-Difference Time-Domain Method for Numerical Modeling of Electromagnetic Wave Interactions with Arbitrary Structures*, Vol. PIER 2, edited by M. A. Morgan, 287–373. New York: Elsevier, 1990.
28. Peterson, A. F., S. L. Ray, and R. Mittra., *Computational Methods for Electromagnetics*. Oxford: Oxford University Press, 1998.
29. Chiprout, E., and M. S. Nakhla., *Asymptomatic Waveform Evaluation and Moment Matching for Interconnect Analysis*. Boston, MA: Kluwer Academic Publishers, 1994.
30. Pasha, S., A. C. Cangellaris, J. L. Prince, and M. Celik. "Passive Model Order Reduction of Multiconductor Interconnects." *Seventh IEEE Topical Meeting on Electrical Performance of Electronic Packaging*, 291–4. West Point, NY, October 1998.
31. Davidson, C. W., *Transmission Lines for Communications*. New York: Wiley, 1978.
32. Ramo, S., and J. R. Whinnery., *Fields and Waves in Modern Radio*. New York: Wiley, 1953.
33. Pasha, S., M. Celik, A. C. Cangellaris, and J. L. Prince. "Passive SPICE Compatible Models of Dispersive Interconnects." In *Proceedings of 49th IEEE Electronic Components and Technology Conference*, 493–9. San Diego, CA, June 1999.
34. Weeks, T. "Calculation of Coefficients of Capacitance of Multiconductor Transmission Lines in the Presence of a Dielectric Interface." *IEEE Trans. Microwave Theory Tech., MTT.* 18, no.1 (1970): 35–43.
35. Ruehli, A. E. "Inductance Calculations in a Complex Integrated Circuit Environment." *IBM J. Res. Dev.* 16, no. 5 (1972): 470–81.
36. Ruehli, A. E. "Survey of Computer-Aided Electrical Analysis of Integrated Circuit Interconnections." *IBM J. Res. Dev.* 23 (1979): 626–39.
37. Vakanas, L., S. Hasan, A. C. Cangellaris, and J. L. Prince. "Effects of Floating Planes in Three-Dimensional Packaging Structures on Simultaneous Switching Noise." *IEEE Trans. Comp., Packag., Manuf. Technol B: Adv. Packag.* 21 (1998).
38. Hasan, S., A. Cangellaris, and J. Prince. "A New RLC Modeling Tool Based on the Partial Element Equivalent Circuit (PEEC) Technique." In *Proceedings of the Fourth VLSI Packaging Workshop of Japan*, 59–64. Kyoto, Japan, November 1998.
39. Vakanas, L. P., A. C. Cangellaris, and J. L. Prince. "Frequency-Dependent [L] and [R] Matrices for Lossy Microstrip Lines." *Trans. Soc. Comput. Sim.* 8 (1991): 281–318.
40. Davidson, E. E. "Electrical Design of a High Speed Computer Package." *IBM J. Res. Dev.* 26, no. 3 (1982): 349.
41. Senthinathan, R., and J. L. Prince. "Simultaneous Switching Ground Noise Calculation for Packaged CMOS Devices." *IEEE J. Solid-State Circuits* 11 (1991): 1724–8.

42. Gribbons, M., A. C. Cangellaris, and J. L. Prince. "Finite-Difference Time-Domain Analysis of Pulse Propagation in MCM Interconnects." *IEEE Trans. Comp., Hybrids, Manuf. Technol.* 16 (1993): 490–8.
43. Ruehli, A. E., and A. C. Cangellaris. "Progress in the Methodologies for the Electrical Modeling of Interconnects and Electronic Packages." *Proc. IEEE* 89, no. 5 (2001): 740–71.
44. Harrington, R., *Field Computation by Moment Methods*. Malabar, FL: Robert E. Krieger Publishing Co., 1968.
45. Palusinski, O. A., J. C. Liao, P. E. Teschan, J. L. Prince, and F. Quintero. "Electrical Modeling of Interconnections in Multilayer Packaging Structures." *IEEE Trans. Comp., Hybrids, Manuf. Technol.* CHMT-10 (1987): 217–23.
46. Omer, A. A., A. C. Cangellaris, M. M. Mechaik, and J. L. Prince. "The Per-Unit-Length Capacitance Matrix of Flaring VLSI Packaging Interconnections." *IEEE Trans. Comp., Hybrids, Manuf. Technol.* 14 (1991): 749–54.
47. Cangellaris, A. C., J. L. Prince, and L. Vakanas. "Frequency-Dependent Inductance and Resistance Calculation for Three-Dimensional Structures in High-Speed Interconnect Systems." *IEEE Trans. Comp., Hybrids, Manuf. Technol.* CHMT-13 (1990): 154–9.
48. Kamon, M., M. J. Tsu, and J. White. "Fasthenry, a Multipole-Accelerated 3-D Inductance Extraction Program." In *Proceedings of ACM/IEEE Design Automation Conference*. Dallas, TX, June 1993.
49. Palusinski, O. A., J. C. Liao, J. L. Prince, and A. C. Cangellaris. "Simulation of Transients in VLSI Packaging Interconnections." *IEEE Trans. Comp., Hybrids, Manuf. Technol.* CHMT-13 (1990): 160–6.
50. Liao, J. C., O. A. Palusinski, and J. L. Prince. "Simulation of Transients in VLSI Packaging Interconnects with Frequency-Dependent Parameters and Nonlinear Terminations." *IEEE Trans. Comp., Hybrids, Manuf. Technol.* CHMT-13 (1990): 833–8.
51. Cooke, B. J., J. L. Prince, and A. C. Cangellaris. "S-Parameter Analysis of Multiconductor Integrated Circuit Interconnect Systems." *IEEE Trans. Comput. Aided Des. Integrated Circuits Syst.* 7 (1992): 353–360.
52. Baumgartner, C., and O. A. Palusinski. "Simulation of Lossy Transmission Lines." *Proc. IEEE Topical Meeting Electrical Perform. Electron. Packag.* April (1992): 161–3.
53. Celik, M., *AZSPICE Circuit Simulator Using Macromodels, Modified from Berkeley Spice3f4*. Tucson: University of Arizona, 1996.
54. Pasha, S., A. C. Cangellaris, J. L. Prince, and M. Celik. "A New Discrete Transmission Line Model for Passive Order Model Reduction and Macromodeling of High Speed Interconnects." *IEEE Trans. Comp., Packag., Manuf. Technol. B: Adv. Packag.* 22, no. 3 (1998): 356–64.
55. Davis, E. M., W. E. Harding, R. S. Schwartz, and J. J. Corning. "IBM." *J. Res. Dev.* 8 (1964): 102.
56. Goldmann, L. S., and P. A. Totta. *Solid State Technology* June (1983).
57. Koopman, N. G., T. C. Reiley, and P. A. Totta. "Chip-to Package Interconnections." In *Microelectronics Packaging Handbook*, edited by R. R. Tummala, and E. J. Rymaszewski, 361–458. New York: Van Nostrand Reinhold, 1989, Chap. 6.
58. See discussions in Koopman, N. G., and P. A. Totta. *ASM World Materials Conference*, 1988.
59. Suryanarayana, D., R. Hsiao, T. P. Gall, and J. M. McCreary. *IEEE Trans. Comp. Hybrids, Manuf.* 14 (1991): 218.
60. Powell D. O., and A. K. Trivedi. *Proceedings of the IEEE Electronics Components and Packaging Technology Conferences*, 182, 1993.
61. Dai, X., C. Kim, R. Willecke, and P. S. Ho. *Electron. Packag. Mater. Sci. IX, Mater. Res. Soc. Symp. Proc. J.* 445 (1996): 167.
62. Guo, Y., C. K. Lim, W. T. Chen, and C. G. Woychik. *IBM J. Res. Dev.* 5 (1993): 635.
63. Post, D., B. Han, and P. Ifju. *High Sensitivity Moire: Experimental Analysis for Mechanics and Materials*. New York: Springer, 1994.
64. Dai, X., C. Kim, R. Willecke, S. W. Poon, and P. S. Ho. *Exp./Numer. Mech. Electron. Packag.* 1 (1996): 15.
65. Guo, Y., D. Post, and R. Czarnek. *Exp. Mech.* 29, no. 2, (1989).

66. Choi, S. Y. M. S. "Curing and thermal stress of underfill liquid encapsulants for microelectronics packaging". thesis, The University of Texas, 1997.
67. Dai, X., J.-H. Zhao, and P. S. Ho. "Thermal Stress and Strain in Flip-Chip Packaging-Moiré Interferometry Measurement and Finite Element Analysis". In *Experimental/Numerical Mechanics in Electronic Packaging*, edited by R. Mahajan et al., Vol. 2, 55–63. Society for Experimental Mechanics, 1997.
68. Suhir, E. *Proceedings of the 37th Electronics Components Conferences, IEEE*, 508, 1987.
69. Guo, Y., W. T. Chen, and C. K. Lim. *Advances in Electronic Packaging, Proceedings of the ASME Conference*, 779. San Jose, CA, 1992.
70. Han, B., and Y. Guo. *J. Electron. Packag., ASME* 117, (1995).
71. Zou, D., X. He, S. Liu, Y. Guo, and F. Dai. *Appl. Exp. Mech. Electron. Packag., ASME* EEP-22 (1997): 69.
72. Peterson, D. W., J. N. Sweet, S. N. Burchett, and A. Hsia. *IEEE Electron. Comp. Tech.* (1997).
73. Goodelle J., R. A. Pearson, and T. Y. Wu. *ASME Symposium on Application of Fracture Mechcanics in Electronic Packaging*. San Francisco, CA, 1996.
74. Wu, T. Y., and G. H. Thiel. *ASME Symp. Appl. Fract. Mech. Electron. Packag.* EEP-11 (1994): 205.
75. Jiao, J., C. Gurumurthy, E. Kramer, Y. Sha, C. Y. Hui, and P. Borgensen. *ASME Symp. Appl. Fract. Mech. Electron. Packag.* AMD-222/EEP-20 (1997): 97.
76. Dai, X., M. V. Brillhart, and P. S. Ho. *ASME Symp. Appl. Fract. Mech. Electron. Packag.* AMD-222/EEP-20 (1997): 115.
77. Wang, J., M. Lu, D. Zou, and S. Liu. *ASME Symp. Appl. Fract. Mech. Electron. Packag.* AMD-222/EEP-20 (1997): 103.
78. Chen, W. T., D. Questad, D. Read, and B. Sammakia. *ASME Symp. Appl. Fract. Mech. Electron. Packag.* AMD-222/EEP-20 (1997): 183.
79. Chen, W. T., and C. W. Nelson. *IBM J. Res. Dev.* 23 (1979): 179.
80. Pao, Y. H., and E. Eisele. *J. Electron. Packag.* 113 (1991): 164.
81. Mirman, S. B. *J. Electron. Packag.* 114 (1992): 384.
82. Iannuzzelli, R. J., J. M. Pitarresi, and V. Prakash. "Solder Joint Reliability Prediction by the Integrated Matrix Creep Method." *J. Electron. Packag.* 118 (1996): 55–61.
83. Syed, A. "Factors Affecting Creep-Fatigue Interaction in Eutectic Sn/Pb Solder Joints." *Advances in Electronic Packaging*. EEP-Vol. 19-2, 1535–8. New York: ASME, 1997.
84. Darveaux, R. "How to Use Finite Element Analysis to Predict Solder Joint Fatigue Life." In *Proceedings, 6th International Congress on Experimental Mechanics*. 41–8. New York: Elsevier, 1996.
85. Darveaux, R. "Solder Joint Fatigue Life Model." *Design and Reliability of Solder and Solder Interconnections, Proceedings of the TMS*, 213–8. Warrendale, PA: The Minerals, Metals, and Materials Society, 1997.
86. Kitano, M., A. Nishimura, and S. Kawai. "Analysis of Package Cracking during Reflow Soldering Process." In *Proceedings of International Reliability Physics Symposium*, 90–5. 1988.
87. Madenci, E., I. Guven, and B. Kilic., *Fatigue Life Prediction of Solder Joints in Electronic Packages with ANSYS*. Boston: Kluwer Academic Publishers, 2002.
88. Wong, E. H., Y. C. Teo, and T. B. Lim. "Moisture Diffusion and Vapour Pressure Modeing of IC Packaging." In *Proceedings of 48th Electronic Components and Technology Conference*, 1372–8. 1998.
89. Wong, E. H., K. C. Chan, R. Rajoo, and T. B. Lim. "The Mechanics and Impact of Hygroscopic Swelling of Polymeric Materials in Electronic Packaging." In *Proceedings of 50th Electronic Components and Technology Conference*, 576–80. Las Vegas, NV, May 2000.
90. Wong, E. H., S. W. Koh, K. H. Lee, and K. M. Lim. "Advances in Vapour Pressure Modeling in Electronic Packaging." *Third International Conference on Benefiting from Thermal and Mechanical Simulation in (Micro-)Electronics, EuroSIME*, 347–55. Paris, 2002.

33

300 mm Wafer Fab Logistics and Automated Material Handling Systems

33.1 Introduction **33**-2

33.2 Wafer and Reticle Metrics, Carriers, and Tracking Systems **33**-3
Purpose • Wafer Handling Metrics and Identification • Reticle Handling Metrics and Identification • Elements of Carriers and Tracking Systems

33.3 Interbay Transport and Storage **33**-10
Purpose and Typical Applications • Interbay System Elements and Configuration • Interbay System Planning • Benefits of Interbay Systems

33.4 Intrabay Transport and Storage **33**-24
Purpose and Typical Applications • Intrabay System Configurations • Major Elements of Automated Intrabay • Implementation Barriers and Guidelines for Overcoming Them • Key Performance Factors for Intrabay Systems • Intrabay System Throughput Constraints and Possible Solutions • Intrabay System Sizing and Planning • 300 mm Process and Support Equipment Interfaces • Buffering on the Production Equipment for Uninterrupted Processing: • Benefits of Intrabay Systems

33.5 Material Control System **33**-50
Purpose and Typical Applications • MCS Elements, Functions and Key Requirements • Typical MCS Configuration • Key MCS Performance Factors

33.6 AMHS Reliability and Maintainability Requirements **33**-54

33.7 Anomaly Handling **33**-54

33.8 Use of Computer Simulation for Designing and Operating AMHS **33**-55
Benefits of Computer Simulation for AMHS Design and Operation • Typical Computer Simulation Components • How Simulation Is Used

33.9 AMHS Implementation and Related Considerations **33**-57
Implementation Phases and Responsibilities • Facilities Interfaces • Guidance for Installation, Testing, and Training

33.10 Extendibility and Scalability of AMH Systems **33**-64

Appendix **33**-64

References **33**-66

Further Reading **33**-66

Leonard Foster
Texas Instruments, Inc.

Devadas Pillai
Intel Corporation

33.1 Introduction

The purpose of this chapter is to provide technical background and reference information for guidance in planning, specification and application of wafer logistics and automated material handling systems (AMHS) for 300 mm diameter wafer fabrication factories (fabs). 200 mm AMHS related information was presented in the 1st edition of this handbook published by Marcel Dekker, Inc. [1]. Automated material handling system elements, metrics, example implementations, and benefits are summarized. Enabling infrastructures and systems interfaces for 300 mm material handling have also been stressed, where applicable. Illustrations and tables are used to present the concepts and information in a compact, easy to reference format.

Descriptions are given for the elements of interbay and intrabay AMHS and related wafer logistics infrastructure. Interbay refers to the movement of wafers between bays or areas of process tools in the fab while intrabay refers to the movement of wafers between process tools within a given bay of the fab. The material handling equipment elements of interbay AMHS include stockers or automated storage and retrieval systems (AS/RS) for storing wafers in carriers and a choice of three types of carrier horizontal transport between stockers which include: (1) ceiling supported, overhead monorail tracks with transport vehicles; (2) floor based, guided vehicles; and (3) cleanroom qualified, powered roller conveyors. In addition, automated "vertical lifters" are used to transport carriers vertically between interbay stockers or transport equipment on different floor levels of the fab. The floor-based types of intrabay material handling equipment that can be used include: (1) automated guided vehicle (AGV) and (2) rail guided vehicle (RGV); and the ceiling-supported types of intrabay material handling equipment include: (1) the overhead hoist vehicle (OHV) transport (OHT) and (2) the powered roller conveyor with vertical lifters. In order to operate as a system, the interbay and/or intrabay equipment require a highly reliable material control system (MCS) to direct the operation of the equipment in the routing of wafer carriers in the fab in accordance with product flow information from the manufacturing execution system (MES) or factory level control system. For effective intrabay operation, an automated carrier dispatch scheduling system is also required. From an AMHS trend and roadmap perspective, OHV transport has been the predominate means of transport in the first few generations of 300 mm fabs; however, it is expected that powered roller conveyors will be used more in the future due to increases in wafer carrier throughput requirements. Floor-based transportation vehicles (AGVs, RGVs) have seen less application in 300 mm lines due to throughput requirements and safety concerns co-existing with fab operations personnel.

A bibliography of wafer fab logistics and AMHS related resource materials is provided along with a list of key Semiconductor Equipment and Materials International (SEMI) industry standards [2] in the Appendix. For more updated information, the reader is referred to the Web sites of SEMI www.semi.org for SEMI standards and activities related to the physical interfaces and carriers (PIC) Committee and the Information and Controls Committee; to International SEMATECH www.sematech.org for 300 mm semiconductor factory integration guidelines [3–5]; and to the current edition of the International Technology Roadmap for Semiconductors (ITRS) www.itrs.net [6] for current and future wafer fab logistics requirements.

The goals of wafer fab logistics and AMHS are to cost efficiently move wafers and reticles between the various units of fab production equipment in support of the overall manufacturing cycle time, throughput, and operational requirements of the fab. Key considerations for efficient wafer fab logistics include unique identification and accurate reading of wafers, reticles, and their carriers; reliable, fast operation of material handling equipment without impact to wafers or device yields; reduction of floor space requirements for wafer "work-in-progress" (WIP) storage and transport; compliance with personnel and building safety regulations; and low capital equipment and operating costs. The integration of wafer and reticle carriers, AMHS equipment, and controls systems into productive logistics systems is the challenging work of factory planning and automation engineering staffs.

33.2 Wafer and Reticle Metrics, Carriers, and Tracking Systems

33.2.1 Purpose

The size, weight, and quantity of wafers transported in lot groupings and individually in the wafer fab, dictate the design of wafer carriers and material handling equipment, in addition to shaping the wafer production equipment in which they are processed. In addition, for semiconductor wafers and photolithography masks or reticles to be transported and tracked in the "re-entrant" wafer fab manufacturing line, where there are several hundred manufacturing process steps, they are marked for identification, placed in "clean" carriers, and the carriers are identified for tracking during clean transport and storage.

33.2.2 Wafer Handling Metrics and Identification

The size and weight of the silicon wafers, upon which semiconductor devices are fabricated, determine the sizes and weights of payloads for transport and storage in the fab. A comparison of nominal dimensions and weights of 200 and 300 mm diameter silicon wafers are listed in the Table 33.1.

Wafer identification marking is done by the wafer supplier to support tracking and traceability of the bare wafers during their fabrication and life cycle. Wafer marking may also be done by the wafer fab upon its entry to the fab manufacturing line so that a fab specific lot number plus wafer number and/or a more readable code can be applied. Descriptions of the most frequently used, standardized codes are presented in Table 33.2.

The 2-D matrix code is popular in 300 mm because it enables much smaller wafer edge exclusion capability, resulting in more dies per wafer which is a significant die yield advantage.

33.2.3 Reticle Handling Metrics and Identification

Reticles (or photo masks) are used in photolithographic patterning tools (e.g., steppers and scanners) to define the device pattern for a given process level of the integrated circuit (IC) device. The most commonly used reticles in 300 mm wafer fabs are 150 mm reticles made of quartz. A description of 150 mm reticle metrics and their method of identification are summarized in Table 33.3.

33.2.4 Elements of Carriers and Tracking Systems

Twenty-five wafer capacity, front opening unified pods (FOUPs) are used to transport wafers' lots from one production machine (tool) to the next in the manufacturing flow. Wafers are accessed through the pod door on the front side of the FOUP. The door, interfaces with an automated door opener on each tool load port according to SEMI E62 front opening interface mechanical standard (FIMS). The 25 wafer capacity, front opening shipping box (FOSB) is used to transport wafers from wafer suppliers to wafer fabs and between wafer fab facilities. Front opening shipping boxs can be returned by the fabs to wafer suppliers for reuse. Front opening shipping boxs can be obtained with manually operated doors or doors that interface with tool FIMS ports for automated opening and closing at tool load ports. Generally 150 mm standard mechanical interface (SMIF) pods (single reticle SMIF pod; RSP or multiple reticle

TABLE 33.1 Comparison Metrics for Silicon Wafers

Wafer Metrics (Nominal)	200 mm Wafers	300 mm Wafers
Wafer thickness at center (μm)	725	775
Single wafer weight (gm)	53	128
Weight of 13 wafers (gm)	689	1664
Weight of 25 wafers (gm)	1325	3200
Applicable SEMI standards	M1	M1.15; M 28

TABLE 33.2 Commonly Used Wafer Identification Systems and Related Semi Standards

Wafer ID Systems	ID Location on Wafer	Method of Application and ID Reading	Reference Standard/Description
Alpha-numeric	Wafer frontside or backside, and typically adjacent to wafer notch	Laser marking; human and/or machine vision system reading	SEMI M12 standard for wafer frontside marking with 12 characters including checksum; SEMI M13 standard for wafer marking with 18 characters including checksum
Bar code	Same as above	Laser marking; bar code scanner or machine vision system reading	SEMI T1 standard for wafer back surface marking with bar code; Code 39, and code 412 type bar codes are used for wafers
Two dimensional matrix code (2-D binary data) for 300 mm wafers	Same as above	Laser marking; machine vision system reading	SEMI T7 standard for back surface marking of double-side polished wafers of silicon which comply with SEMI M28 and other materials with diameters of 300 mm and larger

SMIF pod; MRSP) are used to transport photolithography reticles between reticle stockers and photolithography patterning tools within the cleanroom(s) of the wafer fabs. Examples of 25 wafer capacity, 300 mm FOUPs are shown in Figure 33.1 and Figure 33.2 Examples of 25 wafer capacity, 300 mm FOSBs are shown in Figure 33.3 and Figure 33.4 while examples of 150 mm reticle SMIF pods are shown in Figure 33.5 and Figure 33.6. Descriptions of these carriers are provided in Table 33.4.

Wafer and reticle carriers must have certain physical, material, and environmental related attributes to successfully fulfill their functions of holding, transporting, and protecting their contents. Physical attributes are needed for carriers to mechanically engage with and register or "nest" accurately in material handling equipment and at the load ports of production equipment. The environmental design aspects of carriers are interrelated with the requirements of the contents of the carrier, the cleanroom environment, and the equipment of the fab. The physical interface attributes of various commonly used carriers are summarized in Table 33.5. Example manual handles and top robotic flange of the FOUP are shown in Figure 33.7. Example kinematic coupling grooves and gas purge ports at the base of the FOUP are shown

TABLE 33.3 Metrics and Identification for Quartz Photolithography Reticles

Descriptions of 150-mm Reticles	Metrics	Reference Standards
Size of single reticle	152 ± 0.4 mm square ($L \times W$) × 6.22–6.48 mm (H)	SEMI P1 for physical descriptions; easements for ID marks per SEMI P37; alpha-numeric and bar code ID mark location(s) are negotiated between user and suppliers. Data matrix ID mark is defined per SEMI T11
Size of reticle and single pellicle	Pellicle frame size, shape, and height vary according to scanner supplier and user specifications; example pellicle height = 5.9 to 6.3 mm	
Weight of single reticle, approximate	303–390 gm	
Weight of single reticle with single pellicle, approximate	325–415 gm	
Weight of 6 reticles and pellicles, approximate	2300–2500 gm	

FIGURE 33.1 Example 300 mm wafer carrier front opening unified pod (FOUP). (Photograph courtesy of Shin-Etsu Polymer Co., Ltd.)

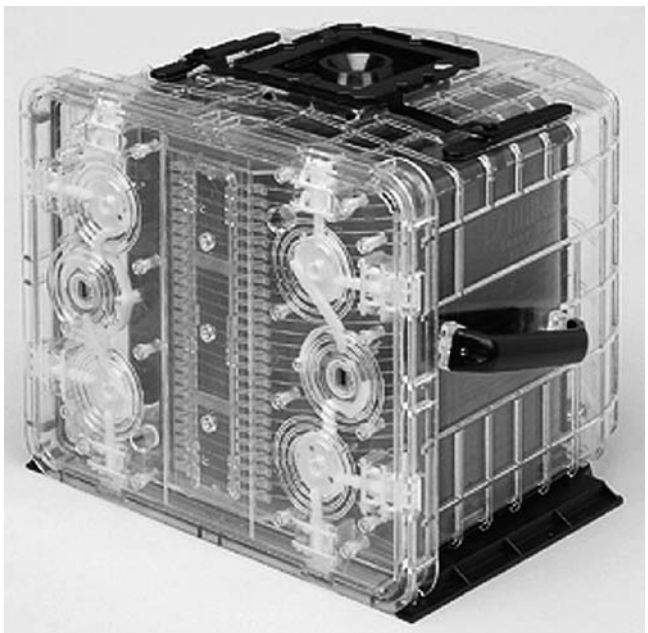


FIGURE 33.2 Example 300 mm wafer carrier (FOUP). (Photograph courtesy of Dainichi Shoji K.K.)

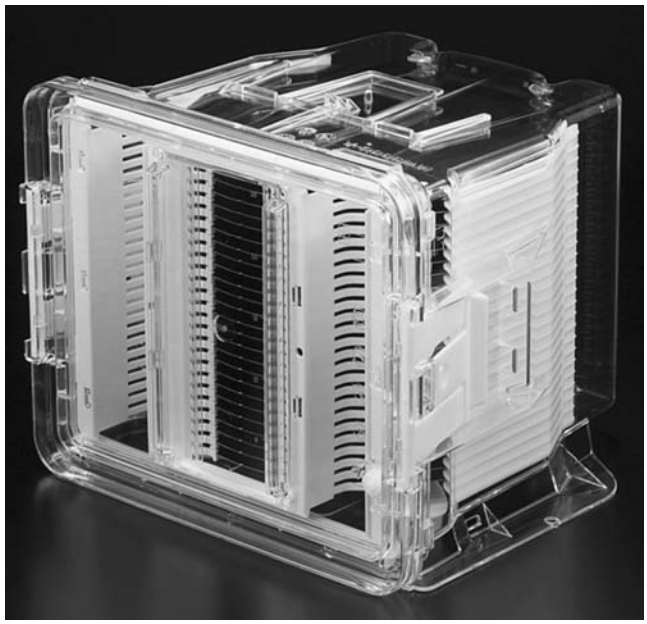


FIGURE 33.3 Example 300 mm front opening shipping box (FOSB) with manual door. (Photograph courtesy of Shin-Etsu Polymer Co., Ltd.)

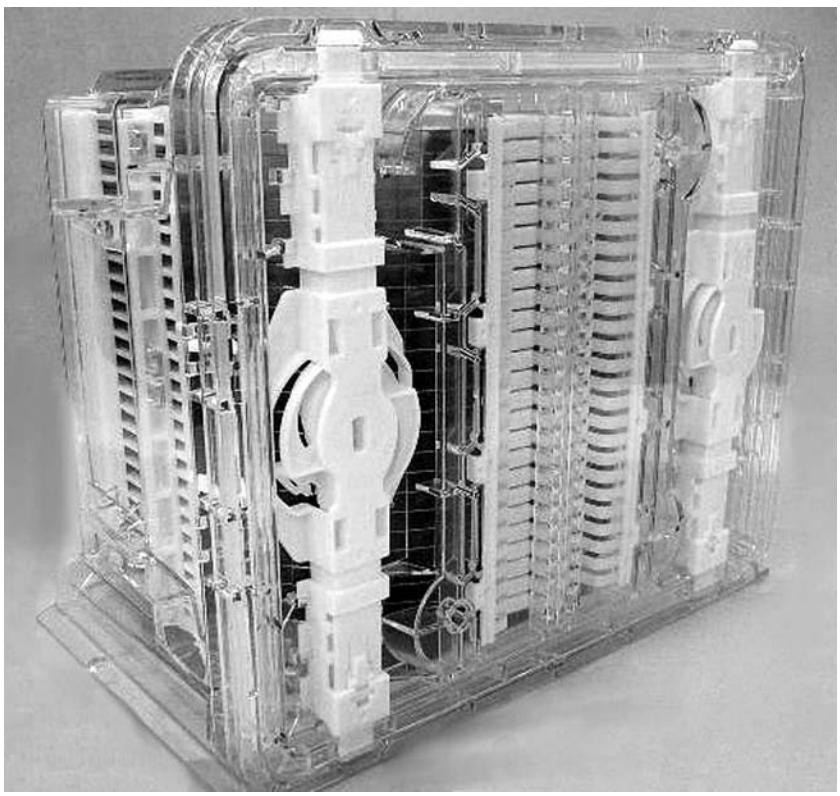


FIGURE 33.4 Example 300 mm FOSB. (Photograph courtesy of Miraial Co., Ltd.)

FIGURE 33.5 Example 150 mm RSP. (Photograph courtesy of Dainichi Shoji K.K.)

FIGURE 33.6 Example 150 mm RSP. (Photograph courtesy of Entegris, Inc.)

TABLE 33.4 Descriptions of Commonly Used Carriers and Typical Payloads in 300 mm Fabs

Wafer Carrier Descriptions	Application and SEMI Standard (Carrier Capacity)	Carrier Feature Description	Carrier Materials	Size $W \times D \times H$ (mm)	Environmental Features	Weights of Carrier, Contents, and Total Payload (kg)
Front opening unified pod (FOUP)	Wafer intra-factory transport per E47.1 (13 or 25 wafers)	Wafer support teeth at 10 mm pitch are integrated into pod shell. Has front opening removable door and kinematic coupling grooves at bottom	Pod shell and door are typically polycarbonate; Structural members are generally carbon fiber filled, engineering polymers	430×356×338 for 25 wafer pod with manual handles and top robotic handling flange; 430×356×215 for 13 wafer pod with same features	Semi-isolated mini-environment interior; inert gas purging is possible option with two and four purge port models available among carrier suppliers	Carriers: ~3.2 for 13 capacity and ~4.5 for 25 capacity; Total payload: ~4.9 for 13 wafers and 7.8–8.0 for 25 wafers
Front opening shipping box (FOSB) for 300 mm wafers	Wafer inter-factory transport per M31 (25 ea. Wafers) or E119 Reduced Pitch Pod (25 ea. Wafers)	See above. M31 box has 10 mm wafer pitch. E119 box has 6.25 mm wafer pitch spacing	Polycarbonate shell and door; thermoplastic elastomer wafer retainers	430×356×339 for 25 wafer full pitch pod; 430×356×228 for 25 wafer reduced pitch pod; both with manual handles and top robotic handling flange	Semi-isolated mini-environment interior; carrier generally wrapped and sealed with plastic film when used as wafer shipper between factory sites	Carrier: 4.6–4.8; Total payload: 7.8–8.0 for 25 wafers
Standard mechanical interface (SMIF) pods [bottom opening]: reticle SMIF Pod and multiple reticle SMIF pod	Reticle intra-factory transport per E111 (1 ea. 150 mm reticle—RSP) or E112 (6 ea. 150-mm reticles—MRSP)	Cassette is distinct and separate from the pod shell and the bottom opening door	Pod shell is polycarbonate; cassette is usually a metal structure with non-metallic inserts for reticle support	RSP=215×230.5×78.5; MRSP=215×230.5×208.5	Semi-isolated mini-environment interior; inert gas purging is possible option	Carriers: RSP=~1; MRSP=1.5; Total payload: RSP=~1.4; MRSP=~4.0

TABLE 33.5 Carrier Types and Physical Interfaces

Carrier Types (Per SEMI Standards)	Equipment Interface Features (Per SEMI Standards)	Features for Manual Handling	Features for Automated Handling	ID Method for Tracking
Standard mechanical interface (SMIF) pod for 150 mm reticles (E111 and E112)	Auto-removable bottom door, rounded corners, registration pin holes, and door latch mechanisms (E19.3–150 mm SMIF); interior cassette (1 or 6 reticle capacity)	Handles or flange attachable to pod shell	Mushroom shaped flange (per SEMI E47.1 and 142 mm square) at top of pod shell	Printed label or electronic ID tag
Front opening, unified pod (FOUP) for 300 mm wafers (E47.1)	Kinematic coupling of three grooves at 120 degree intervals at base of pod; auto-removable front side door compatible with tool front opening interface mechanical standard (FIMS) ports (E15.1, E57, E62); purge gas ports at base per supplier locations	Power grip handles at two sides of pod	Mushroom shaped flange (142 mm square) at top of pod shell; secondary kinematic coupling grooves and outboard conveyor rails on the bottom	Printed label or electronic ID tag
Front opening shipping box (FOSB) for 300 mm Wafers (M31) or Reduced pitch box (E119)	Kinematic coupling of three grooves at 120 degree intervals at base of pod and auto-removable door or manual door (E15.1, E57, E62)	Power grip handles	Mushroom shaped flange at top of pod shell; secondary kinematic coupling grooves and outboard conveyor rails on the bottom	Printed label or electronic ID Tag

FIGURE 33.7 Back side of FOUP showing top robotic flange, manual handles, and identification tag holder. (Photograph courtesy of Entegris, Inc.)

in Figure 33.8. In addition, carriers must be qualified for wafer fab use with respect to their materials of construction, structural stability, and wear resistance properties. Extensive testing of carrier materials and assessment of trace elements within the carrier materials should be done prior to their use in a fab; however, details on carrier materials testing is beyond the scope of this chapter.

An identification (ID) system is employed to give each carrier a unique ID, and it is through this ID system that wafer and reticle carriers are tracked as they are moved between factories and within the fab, by both manual and AMHS. These ID systems support human and/or machine readable labels or tags. Fab operations personnel require human readable, alpha-numeric ID labels and AMHS equipment require machine readable labels or tags. The popular bar code label and radio frequency identification (RF ID) versions of carrier ID systems are described in Table 33.6.

33.3 Interbay Transport and Storage

33.3.1 Purpose and Typical Applications

The purpose of a 300 mm fab interbay AMHS is to store and transport wafers or reticles in carriers between storage and vertical lift equipment in the various processing areas or bays of the fab. With the appropriate carrier mechanical interfaces, the interbay equipment described herein can be applied to wafer and reticle handling either as separate systems or within the same system of equipment. A MCS inventories the wafer and/or reticle carriers upon their entry to the interbay system and directs the routing of the carriers individually through the AMHS equipment. The routing of the carriers is based on carrier identification and destination information from the MES on the factory control computer.

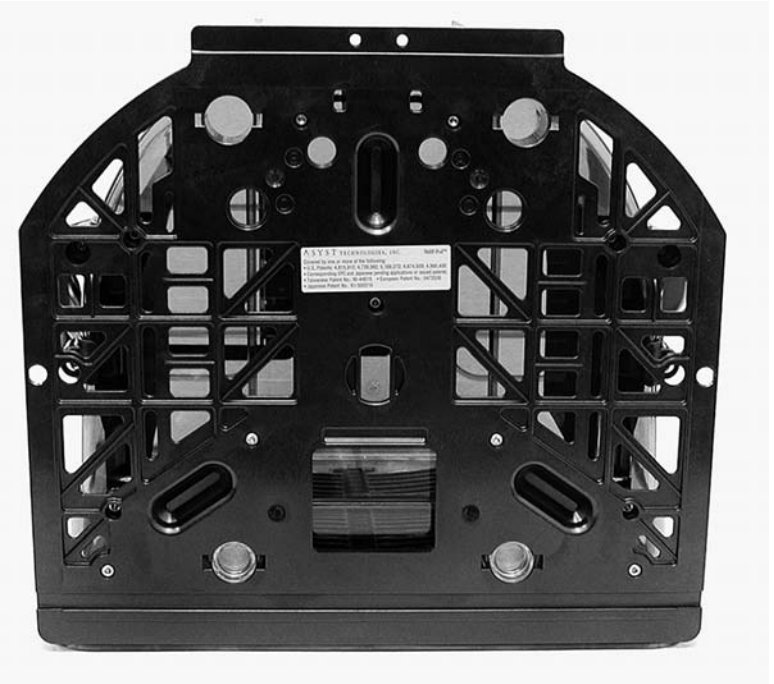


FIGURE 33.8 Bottom of FOUP showing kinematic grooves and purge gas ports. (Photograph courtesy of Entegris, Inc.)

33.3.2 Interbay System Elements and Configuration

33.3.2.1 Major Elements of Interbay Systems

An interbay AMHS may include various quantities and arrangements of the following elements:

- *Stockers or AS/RS*. These programmable, computer controlled systems serve to "pick and place" wafer or reticle carriers to and from a multiplicity of vertically stacked storage locations (or bins) from and to input/output ports which hand off carriers to AMHS transport elements or fab operating personnel.
- *Transport systems*. These are used to transport carriers between processing areas or bays in a factory. Usually the transport systems are overhead monorail tracked vehicle type systems. Floor running AGV systems are used in applications where overhead transport is not practical. Generally, the overhead monorail system has much higher throughput capability compared to the AGV system and it requires less floor space.
- *Inter-level lift system*. If processing is done in multiple levels of a cleanroom, then inter-level lift systems are used to transport carriers between these different levels.
- *Carrier identification systems*. These systems are responsible for reading the identification of carriers as it passes through the different elements of the material handling system. The reading is generally done at the input ports of the stockers and lift systems; however, it may also be done at the output ports for increased traceability.
- *Control systems*. Equipment level and intermediate host based control systems coordinate and manage the transport, handling, storage, and identification functions of the AMHS.
- *User interfaces*. These are computer screens and user data input/output units, located in different places in the factory, where production personnel can access and enter information on the status

TABLE 33.6 Carrier Identification and Tracking Systems Summary

Carrier ID Systems	Base Technology	Features	Label/Tag Use Locations	User Interfaces	Advantages	Disadvantages
Bar code (BC) label	Bar code laser scanner, class 2	E.g., Code 39 bar code with check sum	Tops and sides of carriers	Machine readable with electronic display or terminal for human read	BC readers are compact and reliable	Bar code by itself is not human readable and requires a clean printing method. No on-board memory. Mounting of the reader in the equipment may be difficult
Radio frequency identification (RF ID) transponder tag	RF transponder local to payload	Tags have limited ROM or read/write memory capability	Embedded in structure of non-metallic carrier or on exterior of metal carrier	Machine readable with electronic display or terminal for human read	Tag transponders are compact and provide on-board ROM or read/write memory capabilities	Transmission range of transponders limits their use to equipment load ports. RF interference may be a concern at a few tools

of wafer lots/carriers as well as get the operational status of the different components of the material handling system.

33.3.2.2 Interbay Configuration Considerations

Typical layouts of 300 mm interbay AMHS are shown in Figure 33.9. These common interbay configurations are called "spine layouts." In such a layout, a central interbay transport loop (or loops) is installed in the main central aisle of the cleanroom. Stockers located near the entrances of the process bays are connected to the interbay transport system, which consists of monorail track(s) and vehicles (ceiling supported) or AGVs (floor based).

In 300 mm fabs, one or more loops of overhead transport track are most typically used to transport wafer FOUPs or reticle pods between stockers located in the different process bays and cleanrooms of the fab. Intrabay AMHS track systems within the process bays transport the carriers between stocker intrabay I/O ports and the production equipment load ports as directed by the factory's MES and the MCS of the AMHS. Production operators have the option of accessing the carriers at the operator input/output ports of the stockers. Although the layouts in Figure 33.9 show stockers in a cleanroom central aisle only and near the entrances of process bay aisles, some fabs utilize stockers placed in the center of each process bay, or even in the far end of each bay as depicted in Figure 33.10. Each configuration has its "pros and cons" and must first be evaluated by the user prior to final selection.

If carriers are to be transported to a different floor level of the fab, a pair of automated inter-level lifts may be used between stockers or transport input/output conveyors on each level of the fab where carrier routing is needed. Even though one inter-level lift may meet the throughput requirements from one level to the other and vice versa, generally two lifts are recommended for system redundancy reasons.

When intrabay transport systems are limited to their respective process areas, the interbay system transports carriers between the stockers and in turn, the intrabay systems of the fab. Overhead shuttle (OHS) vehicles on monorail track have been the most used type of interbay transport system. Figure 33.11 and Figure 33.12 show two example configurations of interbay systems with respect stocker heights and transport track elevations. In Figure 33.12, it can be seen that a higher ceiling height in a portion of the cleanroom allows the interbay transport system to be above the height of the intrabay tracks. With the emerging implementation of OHT, its OHV, and unified OHT track systems, the role of the interbay system is changed to that of a parallel or backup system. In a unified OHT system, tool-to-tool transport of payloads is achieved without routing payloads through stockers. A more detailed description of OHT and unified OHT is presented later in this chapter. In a fab with a unified transport system and an interbay transport system installed at a higher elevation than the unified system, the OHT vehicles can also be assigned by the MCS to carry out the role of interbay stocker to stocker or lifter transport depending on the transport requirements of the fab. Table 33.7 summarizes these interbay configurations along with their roles, advantages, and disadvantages.

33.3.2.3 Key Performance Factors for Interbay Systems

Interbay systems are judged on their ability to deliver wafer lots to their intended destinations reliably and within an expected time of delivery to the requested destination, without impact to fab personnel safety and product yields. Major performance factors to be considered for system selection are summarized in Table 33.8.

33.3.2.4 Key Characteristics and Performance Capabilities of Interbay AMHS

Key characteristics and performance capabilities of the individual elements of interbay AMHS are summarized in Table 33.9 through Table 33.12. It should be noted that interbay AMHS equipment can be configured to handle wafers in FOUPs or reticles in SMIF pods. Stockers are not equipped to handle both types of carriers in the same stocker; however, different transport vehicles on the same transport track can be equipped to handle FOUPs or SMIF pods.

A photograph of an OHS type transport vehicle is shown in Figure 33.13 and a photograph of an inter-level lifter-stocker is shown in Figure 33.14. Figure 33.15 shows an example FOUP stocker with OHT load

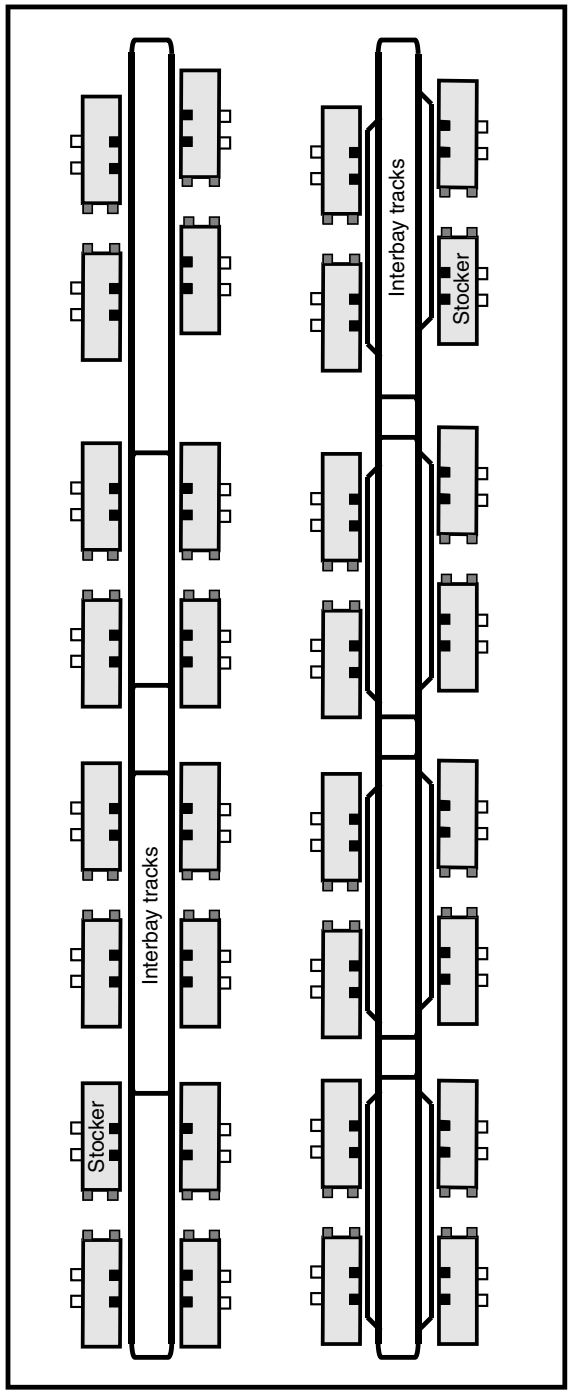


FIGURE 33.9 Two examples of typical interbay storage and transport systems.

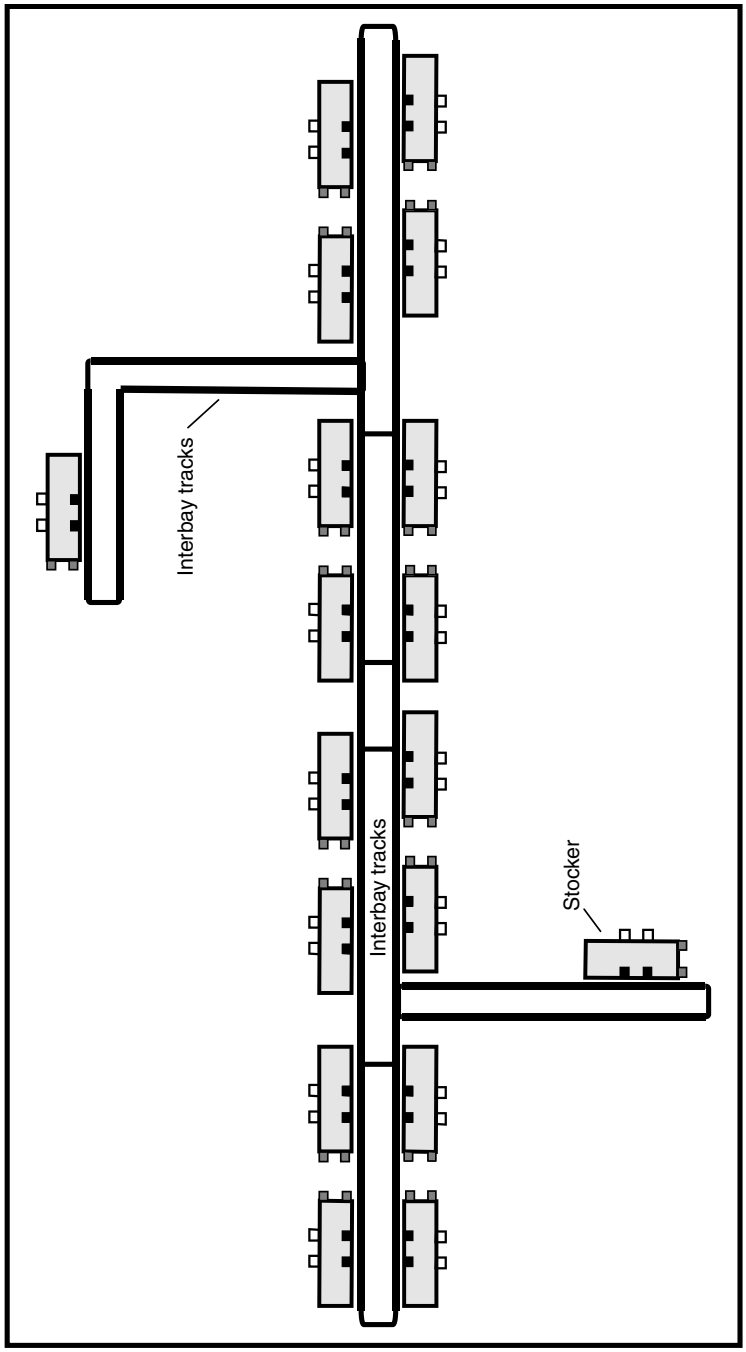


FIGURE 33.10 Alternate interbay layout with some stockers in the middle of bays.

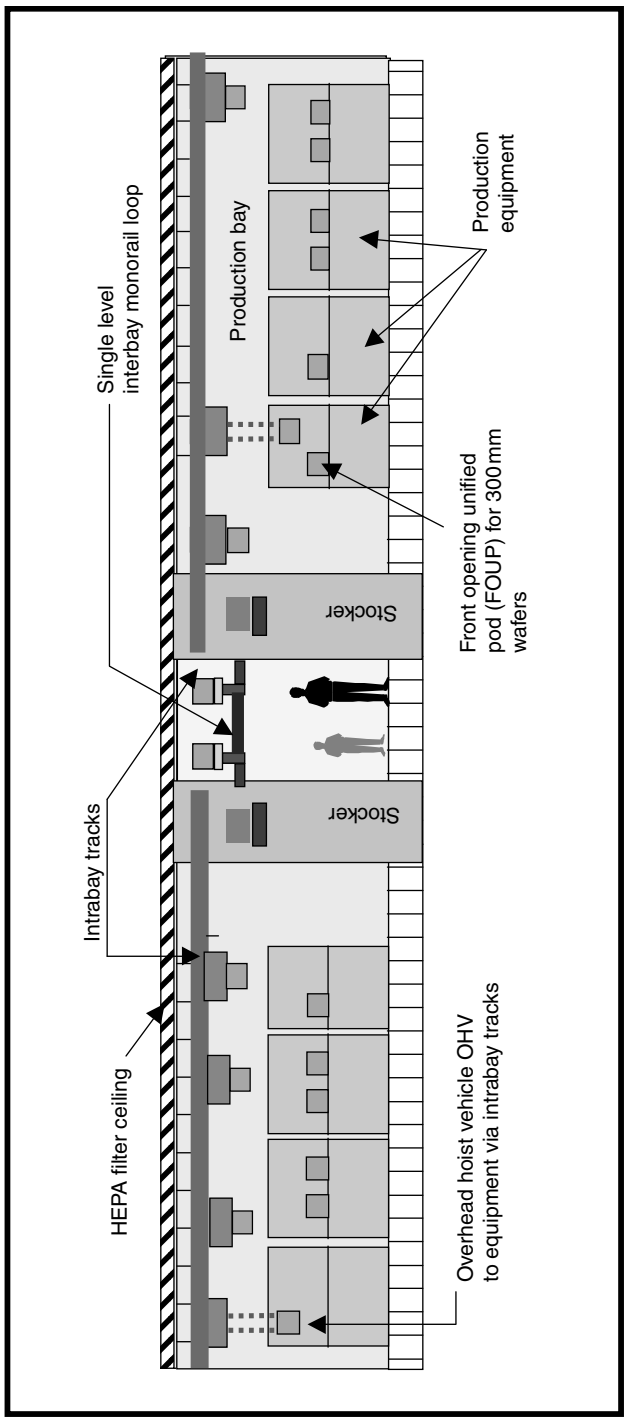


FIGURE 33.11 Cross-section view of a 300 mm fully automated factory showing standard height stockers for storage and single level of interbay transport system. Intrabay track elevation is also represented for reference.

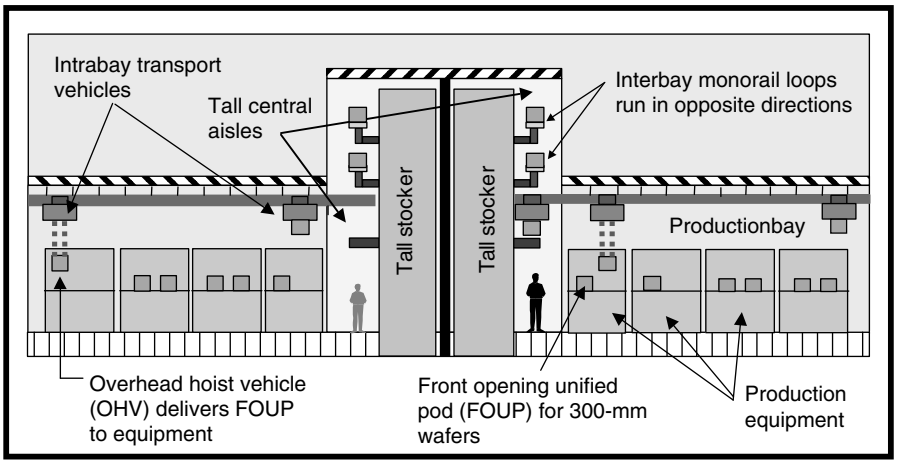

Intrabay transport vehicles
Tall central aisles
Interbay monorail loops run in opposite directions
Tall stocker
Tall stocker
Productionbay
Overhead hoist vehicle (OHV) delivers FOUP to equipment
Front opening unified pod (FOUP) for 300-mm wafers
Production equipment

FIGURE 33.12 Cross-section view of a 300 mm fully automated factory showing tall central aisles for storage and multiple levels of interbay transport systems for higher throughput. Intrabay track elevation also shown for reference.

ports mounted above the floor level and operator load ports mounted on the floor. The open top of a carousel type FOUP stocker, as shown in Figure 33.16, permits automated load and unload of FOUPs by OHT vehicles.

In current 300 mm fabs, reticles are stored in stockers in 150-mm SMIF pods or as bare reticles. When reticles are stored in SMIF pods, they are generally stored in RSP or MRSP. When reticles are stored bare (without a carrier), they are stored in bare reticle stockers, which feature a built-in minienvironment capability and SMIF pod load ports. Key characteristics and performance capabilities of reticle stockers are summarized in Table 33.13. A photograph of an example bare reticle stocker is presented in Figure 33.17.

33.3.3 Interbay System Planning

The application design steps associated with planning and sizing an interbay system are presented in the following flow chart (Figure 33.18).

TABLE 33.7 Interbay Transport Configurations and Their Roles

Configuration	Location	Role	Advantages	Disadvantages
Interbay track separated from intrabay tracks by stockers	Near ceiling, on floor level or in sub-fab	Transport link between process bays and stockers	Implementation is easier	Carriers must travel through stockers resulting in longer carrier delivery times
Interbay vehicles travel on unified intrabay track	Near ceiling and integrated and at same level with intrabay tracks	Interbay transport carried out by the unified transport system	More direct routing of carriers and lower cost of implementation	Higher utilization of unified transport system
Interbay track between stockers and located above unified intrabay track	Near ceiling and above intrabay track or in sub-fab	Transport of lower priority materials and back-up to unified transport system	Lower delivery times and transport redundancy	Cost of the separate interbay transport system

TABLE 33.8 Interbay Performance Factors and Their Importance

Performance Factors	Why It Is Important
System reliability	High reliability in material handling, routing, and tracking is the most important factor for interbay systems and automated material handling systems (AMHS) in general. Interbay systems of medium to large size are expected to have overall uptime of 98% and higher. Refer to SEMI standard. E10 for reliability and uptime calculation methodologies
Stocker cycle time	This is one of the most critical items to be concerned about, because it dictates the overall responsiveness of the AMHS system. Longer cycle time stockers have lower throughput and take longer times to deliver requested wafer lots to the transport vehicles or operators
Interbay transport throughput	Transport throughput capability is a most critical performance metric of the AMHS system. It defines how many carriers or payloads per hour the interbay system is capable of moving across the whole system in a sustained manner. Since the load per hour is directly proportional to the wafer starts per week of the factory, it is very important for users to understand what upside capability the AMHS system has. The AMHS must be designed up front with reasonable upside capability. Otherwise, the interbay system could be the fab bottleneck
Carrier handling quality and cleanliness	Wafer payloads must be handled without damaging or vibrating the wafers and the cleanliness of the carrier handling equipment must be consistent with the requirements of the fab cleanroom(s) in which the handling equipment operates
Storage density of stockers	This is obtained by dividing the overall storage capacity of the stocker by the space of its footprint plus peripheral maintenance space in the cleanroom. The higher the value, the better it is for the factory since the stockers will occupy less cleanroom space. This is a very competitive item for AMHS suppliers, and device manufacturers expect continuous improvement in this metric
Transport track routing capability	This defines the flexibility of the transport track and gives the user an estimation as to how they can route the track to better integrate with various equipment layout arrangements and/or avoid facility obstacles or equipment installation obstacles. Users want the transport routing to be very flexible, so that they can have better opportunity of placing process equipment where they prefer it to be
Transport track turning radius	Another metric for track routing flexibility. Smaller turning radii give users an advantage in routing track in tight areas, and to avoid facility and equipment obstacles
Operator safety and ergonomics	AMHS systems generally co-exist with production operators and maintenance technicians. Therefore it is imperative these systems are equipped with fail-safe operator safety features, which can be invoked, either manually or automatically, where personnel and the electromechanical mechanisms of the interbay system share common space. At payload hand-off points, AMHS equipment, such as stocker load ports must comply with operator ergonomic requirements. See SEMI Standards S2 and E8 for equipment compliance requirements
Inert gas purging	Inert gas purging of front opening unified pod (FOUPs) during storage in stockers permits increased "time windows" between certain wafer process steps where ambient air and/or airborne contaminants impact the topography of the wafer
Fire protection	Since the cassette or the box is made of either polycarbonate or carbon filled polymers, there is a concern of flammability issues when the FOUPs and standard mechanical interface (SMIF) pods are stored in high densities, such as in large stockers. A current methodology is to provide safety showers above the stocker frame, which gets automatically activated through heat or smoke detection systems. Carriers made of fire resistant material may be another way to meet fire protection requirements

33.3.3.1 From/To Matrix

The "From/To" matrix is a 2-D matrix grid, which contains counts of the carrier movements between "sending" and "receiving" units of equipment such as stockers in the case of interbay systems. For example, the matrix can be formed with "From" or sending stockers defining rows of the matrix and with "To" or destination stockers defining the columns of the matrix. Given that stockers are associated with different bays of process tools in the overall fab layout, the movements between pairs of stockers can be counted (and input into the From/To matrix as "1" integer movement) by mentally tracing the path of a given product carrier through its process flow as it moves between process tools, bays, and stockers

TABLE 33.9 Interbay Transport Systems (Ceiling Mounted)

Types	Payload Capacity	Cleanliness Class, Airborne Particles	Typical Transport Speed Range	Typical Track Turning Radii	Typical Track Routing Mechanisms
Overhead shuttle (OHS) monorail and vehicles	1 or 2 carriers, 10–30 kg per vehicle	Class 10 @ ≥0.2 or better	30 m/min in curves to 240 m/min (max on straight track)	0.3–0.5 m	Merge/diverge track; turntable
Overhead hoist transport (OHT) vehicle (OHV)	1 carrier ~ 10 kg per vehicle	Class 10 @ ≥0.2 or better	30 m/min in curves to 180 m/min (max on straight track)	0.3–0.5 m	Merge/diverge track; turntable
Powered roller conveyor	Sequential queuing of carriers using accumulation = /> 10 kg/carrier	Class 10 @ ≥0.2 or better	12–18 m/min adjustable	Right angle turning possible	Powered rollers with carrier travel directions at 90° to each other

TABLE 33.10 Interbay Transport Systems (Floor Based)

Types	Payload Capacity (Max)		Cleanliness Class, Airborne Particles	Transport Speed Range (m/min)	Vehicle Turning Radius
Tape guided automated guided vehicle (AGV)	~20 kg per vehicle	2 Front opening unified pod (FOUPs) or ≥2 standard mechanical interface (SMIF) pods	Class 1 @ ≥0.1 or better	18–60	≥Vehicle length or spin-turn radius, which ever is larger
Free roving AGV	~20 kg per vehicle	2 FOUPs or ≥2 SMIF pods	Class 1 @ ≥ 0.1 or better	18–60	≥ Vehicle length or spin-turn radius, which ever is larger

TABLE 33.11 Interlevel Transport Lift Systems

	Payload Capacity				
Types Lift Equipment	Lifting, Kilograms	Handling	Cleanliness Airborne Particles	Speed Range (m/min)	Throughput Range (Moves/h)
Platform carriage type	10–15	1 carrier per move; configurable carrier storage	Class 10 @ ≥ 0.2 or better	10–15	40–80 depending on travel distance
Tower silo stocker pair	10–15	1 carrier per move; silo height configurable carrier storage	Class 10 @ ≥ 0.2 or better	15–20	50–100 depending on travel distance

as necessary to complete the sequence of semiconductor fabrication process steps. Thus, for a given complement of bay stockers and a given product process flow, the summation of a row of the matrix yields the total carrier movements from a given bay stocker (for one carrier in the flow and independent of the carrier start rate). Similarly, summation of a column associated with a bay stocker yields the total carrier movements into that stocker.

TABLE 33.12 Automated Storage/Retrieval Systems (Stockers) for 25 Wafer Capacity Front Opening Unified Pod (FOUPs)

Stocker, FOUP	Typical Configurations		Size Expandability	Storage Capacity Range	Storage Space Efficiency	Cycle Time
Type/metrics	Robot type	I-O Ports	Height and length, m	FOUPs/stocker	FOUPs/m^2	S
Carousel	Revolving horizontal shelves	Overhead hoist transport (OHT) accessed Shelves at Top	H = 2.4 m to 6.4 m; Fixed 1.77 m $W \times 4.22$ m L	60–250	9–34	0–30
Cartesian	X–Z cartesian + cylindrical arm (5 DOF total)	2 Overhead shuttle (OHS), 2 OHT, 2 operator	H = 3.6–8; L = 2–5; W = 1.5	250–300	23–28	15–18
Silo	Z Tower + cylindrical arm + rotating storage silo (5 DOF total)	2 OHS or 2 OHT, 2 operator	H = 2–4.; L = 1.4–2.5; W = 1.5	30–100	4–10	10–14
Tower	X–Z Tower + cylindrical arm (5 DOF total)	2 OHS, 2 OHT, 2 operator	H = 2–4.5; L = 2–15; W = 1.5	100–200	9–15	12 (100 bin)–16

FIGURE 33.13 Example interbay overhead shuttle (OHS) vehicle. (Photograph courtesy of Asyst Shinko, Inc.)

Summation of the sums of the rows (or columns) yields a measure of the "per carrier per product flow" movements on the interbay transport system. Summation of the row sum and column sum for a given bay stocker yields the "per carrier per product flow" moves, required to be made by the stocker main robot.

33.3.3.2 Movement Matrix Development

1. Associate a stocker with each group of process equipment in the production line.
2. Give a unique identification or number to each stocker in the production line layout.
3. Create a matrix grid (or spreadsheet) with the stocker ID's listed as row headers and column headers.
4. Using the product process flow, mentally route a carrier of wafers through the production line layout from process tool to process tool using the interbay system when needed to move the carrier from a tool in one process bay area to a tool in another process bay area.
5. Each time the interbay system is used to transport the carrier, enter a "1" or increase the count of moves in the appropriate matrix box for the associated "from" stocker and the "to" stocker.
6. Do this inventory of interbay carrier unit moves until the carrier has been "routed" through all the steps of the process flow.
7. Check the symmetry of the sums of rows and columns. The sums of rows and columns associated with a given bay stocker will generally be the same unless there are carrier movements into a process area which are not followed by a complementary move out of the area through "the same stocker."

FIGURE 33.14 Example inter-level lifter-silo stocker. (Photograph courtesy of Daifuku Co., Ltd.)

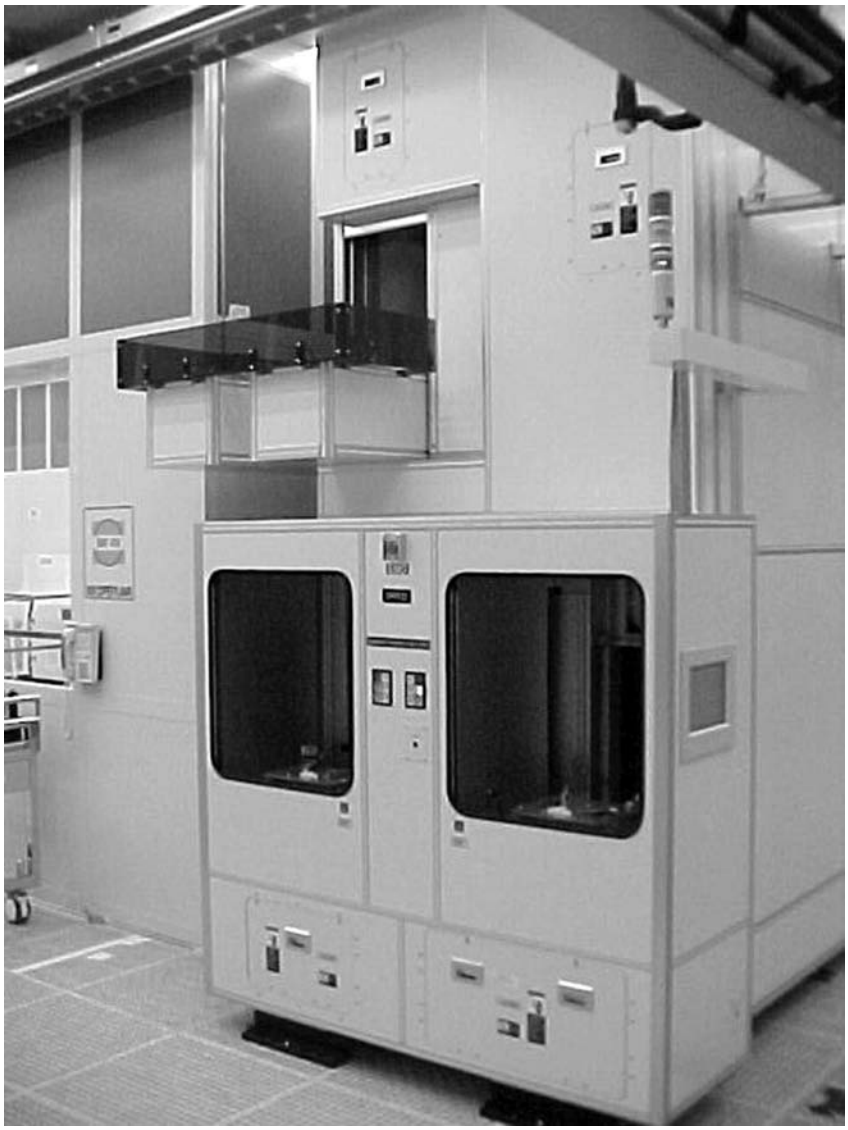


FIGURE 33.15 FOUP stocker operator input/output ports below overhead hoist transport (OHT) ports. (Photograph courtesy of Intel Corporation.)

33.3.3.3 Movement Rates Matrix Development

The "average movement rates" matrix of carrier movements for a given product flow is derived by multiplying the movement matrix by the average rate of carrier starts into the fab production line or interbay system:

$$[\text{Mr}]\text{avg} = [\text{M}]\text{average carrier start rate}$$

An example compilation is shown in Table 33.14.

When mapped into a stocker-to-stocker Movement Matrix the example result is as shown in Table 33.15.

FIGURE 33.16 FOUP carousel stocker top view. (Photograph courtesy of Brooks Automation, Inc.)

33.3.4 Benefits of Interbay Systems

An overview of interbay system benefits is presented in Table 33.16. Quantification of the benefits will vary with each application and are left to the reader. Some financial justification methods can be found in the handbooks on general industrial material handling listed in the bibliography for this chapter.

33.4 Intrabay Transport and Storage

33.4.1 Purpose and Typical Applications

The purpose of intrabay transport systems is to move wafers carriers (FOUPs) and/or reticle SMIF pods (RSPs) between production equipment and stockers in a factory. With the advent of 300 mm high volume manufacturing, where the transport carrier is large and heavy, the adoption of intrabay systems has sky-rocketed. As a result, most 300 mm fabs in the world use it pervasively. The transport of material is accomplished by robotic vehicles that traverse either on the floor or on track near the ceiling. Multiple transport paths will be used by these vehicles depending upon the source and destination of the material being transported. The typical routes for FOUP intrabay are between stocker input/output ports and equipment load ports and from one equipment load port to another equipment load port. During RSP intrabay transport, the transport paths are between reticle stocker input/output ports and the lithography expose equipment RSP load ports, and between lithography expose equipment load ports.

TABLE 33.13 Automated Storage/Retrieval Systems (Stockers) for Reticles

Stocker, Reticle	Typical Configurations		Size Expandability	Storage Capacity Range	Storage Space Efficiency and [Clean Class]	Cycle Time, Robot
Type/metrics	Robot type	I/O ports	Height and length, m	Carriers [Reticles]	Reticles/m^2	S
Single reticle standard mechanical interface (SMIF) pods	*X–Z* Tower + Cylindrical arm (5 DOF total)	2 Overhead shuttle (OHS), 2 overhead hoist transport (OHT), 2 operator	$H = 4.5$; $L = 2–10$; $W = 1.5$	360–1800 [360–1800]	70; [Class 10 @ > 0.2 or better]	15–18
Multiple reticle SMIF pods	*X–Z* Tower + Cylindrical arm (5 DOF total)	2 OHS or 2 OHT, 2 operator	$H = 4.5$; $L = 2–10$; $W = 1.5$	130–650 [780–3900]	150 [Class 10 @ > 0.2 or better]	15–18
Bare reticle storage	*Y–Z* Cartesian arm plus rotating storage silo (5 DOF)	2 OHT and/or 2 operator	$H = 2.9–3.5$; $L = 2.4$; $W = 1.1$	[1300–2000]	200–300 [Class 1 @ > 0.2 or better]	28–32

FIGURE 33.17 Example bare reticle stocker. (Photograph courtesy of Brooks Automation, Inc.)

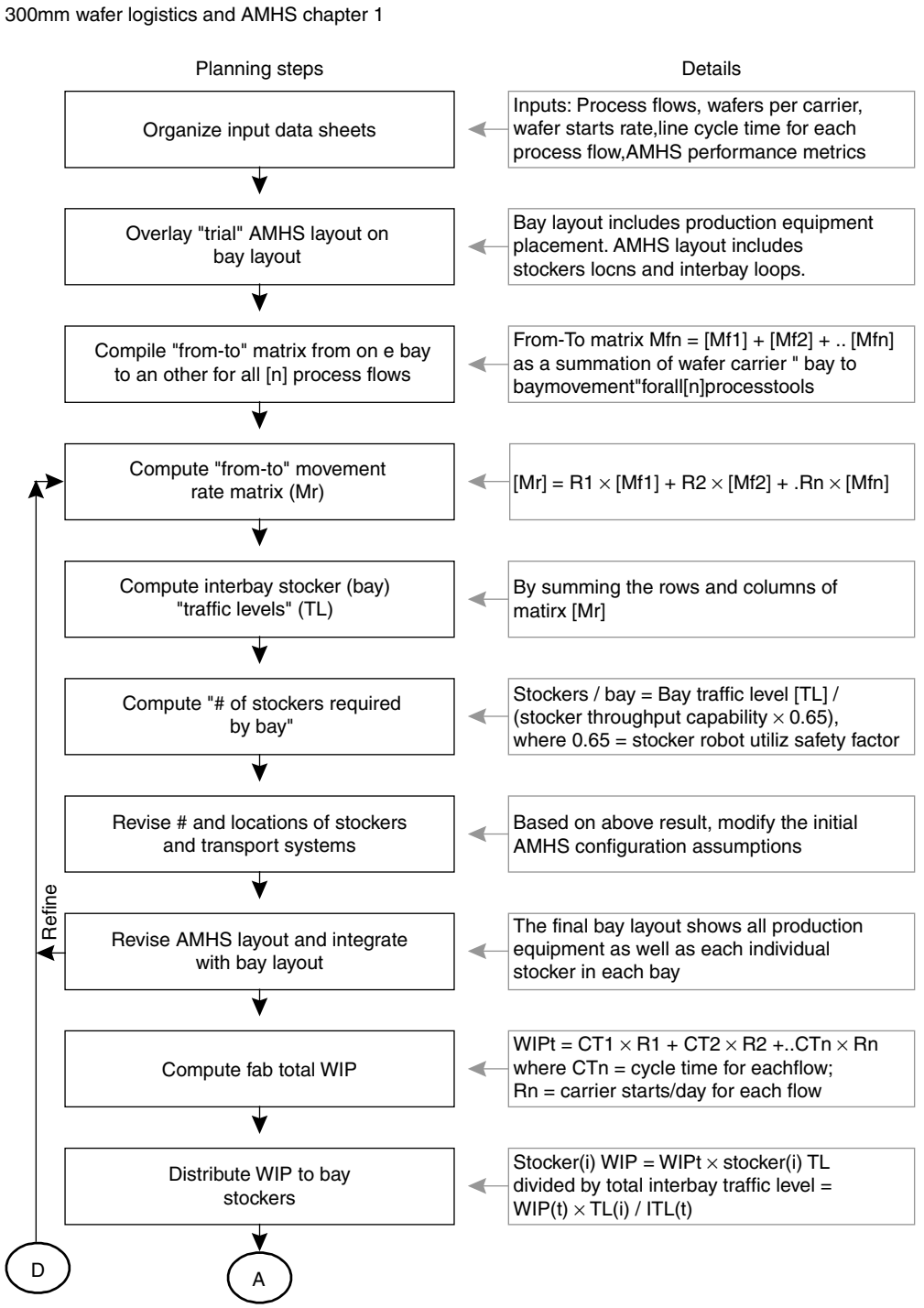

300mm wafer logistics and AMHS chapter 1
Planning steps
Details
Organize input data sheets
Inputs: Process flows, wafers per carrier, wafer starts rate,line cycle time for each process flow,AMHS performance metrics
Overlay "trial" AMHS layout on bay layout
Bay layout includes production equipment placement. AMHS layout includes stockers locns and interbay loops.
Compile "from-to" matrix from on e bay to an other for all [n] process flows
From-To matrix Mfn = [Mf1] + [Mf2] + .. [Mfn] as a summation of wafer carrier " bay to baymovement"forall[n]processtools
Compute "from-to" movement rate matrix (Mr)
[Mr] = R1 × [Mf1] + R2 × [Mf2] + .Rn × [Mfn]
Compute interbay stocker (bay) "traffic levels" (TL)
By summing the rows and columns of matirx [Mr]
Compute "# of stockers required by bay"
Stockers / bay = Bay traffic level [TL] / (stocker throughput capability × 0.65), where 0.65 = stocker robot utiliz safety factor
Revise # and locations of stockers and transport systems
Based on above result, modify the initial AMHS configuration assumptions
Refine
Revise AMHS layout and integrate with bay layout
The final bay layout shows all production equipment as well as each individual stocker in each bay
Compute fab total WIP
WIPt = CT1 × R1 + CT2 × R2 +..CTn × Rn where CTn = cycle time for eachflow; Rn = carrier starts/day for each flow
Distribute WIP to bay stockers
Stocker(i) WIP = WIPt × stocker(i) TL divided by total interbay traffic level = WIP(t) × TL(i) / ITL(t)
D
A

FIGURE 33.18 Interbay system planning flow chart.

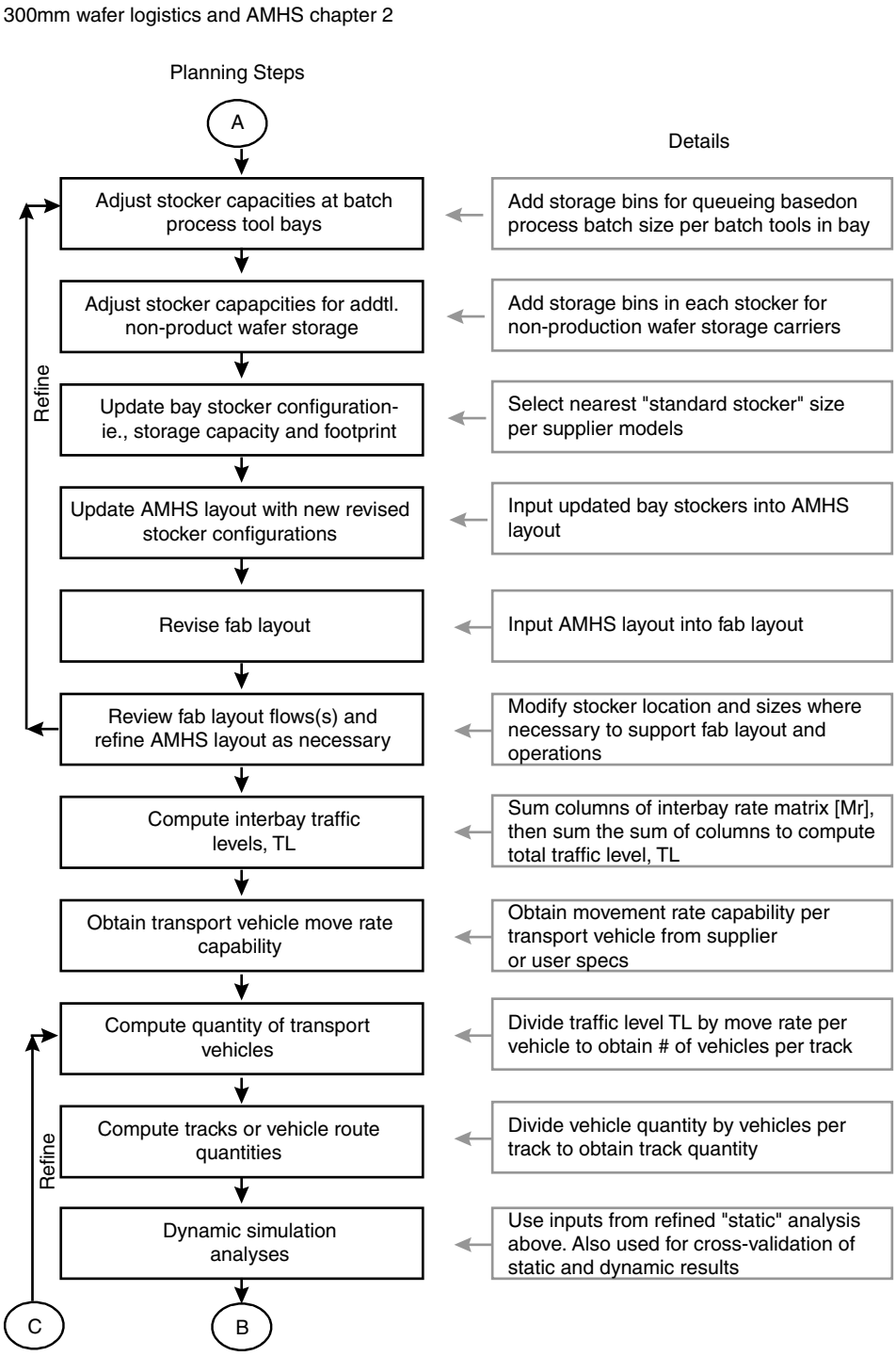

300mm wafer logistics and AMHS chapter 2
Planning Steps
Details
A
Adjust stocker capacities at batch process tool bays
Add storage bins for queueing basedon process batch size per batch tools in bay
Adjust stocker capapcities for addtl. non-product wafer storage
Add storage bins in each stocker for non-production wafer storage carriers
Refine
Update bay stocker configuration- ie., storage capacity and footprint
Select nearest "standard stocker" size per supplier models
Update AMHS layout with new revised stocker configurations
Input updated bay stockers into AMHS layout
Revise fab layout
Input AMHS layout into fab layout
Review fab layout flows(s) and refine AMHS layout as necessary
Modify stocker location and sizes where necessary to support fab layout and operations
Compute interbay traffic levels, TL
Sum columns of interbay rate matrix [Mr], then sum the sum of columns to compute total traffic level, TL
Obtain transport vehicle move rate capability
Obtain movement rate capability per transport vehicle from supplier or user specs
Compute quantity of transport vehicles
Divide traffic level TL by move rate per vehicle to obtain # of vehicles per track
Refine
Compute tracks or vehicle route quantities
Divide vehicle quantity by vehicles per track to obtain track quantity
Dynamic simulation analyses
Use inputs from refined "static" analysis above. Also used for cross-validation of static and dynamic results
C
B

FIGURE 33.18 (*continued*)

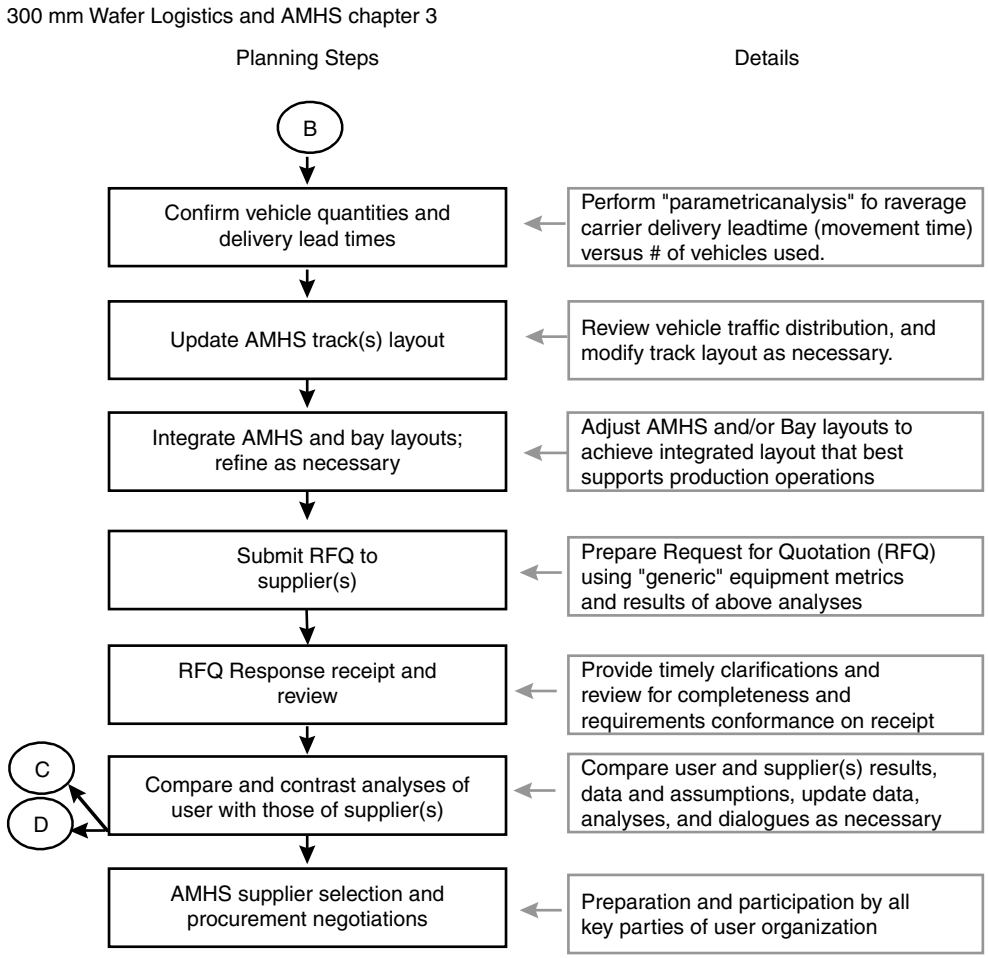

300 mm Wafer Logistics and AMHS chapter 3
Planning Steps
Details
B
Confirm vehicle quantities and delivery lead times
Perform "parametricanalysis" fo raverage carrier delivery leadtime (movement time) versus # of vehicles used.
Update AMHS track(s) layout
Review vehicle traffic distribution, and modify track layout as necessary.
Integrate AMHS and bay layouts; refine as necessary
Adjust AMHS and/or Bay layouts to achieve integrated layout that best supports production operations
Submit RFQ to supplier(s)
Prepare Request for Quotation (RFQ) using "generic" equipment metrics and results of above analyses
RFQ Response receipt and review
Provide timely clarifications and review for completeness and requirements conformance on receipt
C
D
Compare and contrast analyses of user with those of supplier(s)
Compare user and supplier(s) results, data and assumptions, update data, analyses, and dialogues as necessary
AMHS supplier selection and procurement negotiations
Preparation and participation by all key parties of user organization

FIGURE 33.18 (*continued*)

TABLE 33.14 Example Process Flow/Carrier Move Sequence

Step	Bay No. in Which Processing Occurs	Tool No. on Which Processing Occurs	Stocker No. From Which Lot is Taken Out Prior to Processing on Tool	Stocker No. into Which Lot is Placed Back After Processing is Complete on Tool
1	1	Tool 1	1	1
2	4	Tool 3	4	4
3	9	Tool 2	9	9
4	2	Tool 5	2	2
5	3	Tool 1	3	3
6	5	Tool 6	5	4
7	8	Tool 2	8	8
8	4	Tool 3	4	4
9	9	Tool 1	9	9

TABLE 33.15 Example "From/To" Movements Matrix

Stocker No.	1	2	3	4	5	6	7	8	9	Sums
1				1						1
2			1							1
3					1					1
4								1	2	3
5										
6										
7										
8				1						1
9		1								1
Sums		1	1	3	1			1	2	8

33.4.2 Intrabay System Configurations

Modern 300 mm intrabay systems can be configured in two different ways. One configuration is called a unified transport configuration and the other configuration is called a non-unified transport configuration.

In the unified transport configuration, the intrabay transport system tracks and the transport vehicles perform bay to bay material movement (interbay functionality) as well as movement within specific processing bays (intrabay). By this, it means the track network for interbay and intrabay functions are connected directly (without loads having to be ported across these two transport loops using stockers). For this capability to be possible, the track elevations of the two systems must be the same. In such a unified system, once an overhead vehicle picks up a load, then it can transport that load to any other

TABLE 33.16 Benefits of Interbay Systems

Benefit Category	Demonstration Method	Rationale and Comments
Wafer carrier storage efficiency	Ratio of number of carriers per stocker floor space to no. of carriers per manual rack floor space including access space	Since the stocker can be very tall, storage efficiency of stockers is far superior to manual racks for wafers and reticles
Accurate inventory and location of wafers and carriers	Ratio of unassisted carrier average retrieval time to average automated material handling systems (AMHS) assisted retrieval time	Supports cycle time and throughput improvements at constraint tool bays. Tool starvation is minimized
Predictable carrier transport over long distances	Wafer carrier delivery time statistics are logged	Accurate routing of wafers over horizontal and vertical distances
Improved focus on "skill" tasks by fab operating staff	Fab operations staff remain in their assigned process bays and focus on production equipment performance requirements	Reduction of repetitive, unskilled tasks that take operators away from assigned tools and bays. They can focus more or critical process issues (important)
Reduction of ergonomic and safety incidences	Reduced bay-to-bay traffic at floor level	Reduction of repetitive motion injuries as well as lost day cases
Improvements in tool outputs and bay throughput times	Result of operator task simplification and improved focus on production equipment	Constraint tools are less idle since work-in-progress (WIP) is tracked and always at or near these tools. Increased tool output
Reduction in human-induced variability	Wafer carrier movements and tracking are more independent of operator multi-tasking schedules	Carrier delivery independent of human actions which is prone to extensive variations

equipment load port across the entire factory, without the load having to be exchanged through a stocker. Such systems are extremely useful for rapid transport between any two points in the factory. To achieve this, the unified intrabay system has to have significant throughput capability upside and MCS with more advanced vehicle control routing capabilities.

The traditional intrabay AMHS system that uses a separate interbay transport segment (where the load exchange between these two systems occurs through the stockers) is called a non-unified transport system. In such a transport configuration, load movement between different parts of the factory have to be achieved through a combination of the local intrabay vehicle carrying the payload to the tool local bay stocker and then the stocker robot transferring the load to a separate interbay vehicle, which then transports the load to the destination bay stocker. Non-unified transport system can have transport track elevations all at the same level or could be at different levels. A separate intrabay vehicle servicing the destination bay has to pick up the load from the nearby upstream stocker and transport it to the destination tool load port in the bay.

It can be seen that in a unified transport system, only one vehicle (intrabay) is involved in transportation to the destination equipment load port. In the non-unified transport, three sequential vehicle moves are needed (first by the local vehicle to the local stocker, next by the interbay vehicle between stockers and by the intrabay vehicle in the bay of the destination equipment load port). In addition, the stocker robot moves are avoided resulting in additional payload delivery time savings. The advantages of the unified transport configuration are illustrated later in this section.

33.4.3 Major Elements of Automated Intrabay

Major elements of automated intrabay systems include: programmable, computer controlled material handling equipment; carrier identification systems; MCS and related user interfaces; production and stocker equipment load ports; and manual carts to support wafer handling during intrabay AMHS start-up and off nominal conditions. Additional description of the elements of intrabay AMHS are listed as follows:

- *Floor based transport systems*. These include: AGV, and RGV. These were used extensively in 200 mm fabs and are being adapted to a far lesser extent in 300 mm.
- *Ceiling based OHT systems*. These include the OHV. Each vehicle has a hoist mechanism that lowers the carrier to the equipment load port during the load cycle and retrieves the carrier during the unload cycle. This transport technology is being widely proliferated in 300 mm. Examples of OHV transport systems are shown in Figure 33.19 through Figure 33.20.
- *Stocker intrabay input/output ports*. These ports are serviced by the intrabay vehicle, and it is through these ports that the carriers enter and exit the intrabay system. Design and configuration of intrabay input/output ports at the stockers are most critical for high throughput intrabay systems. Generally they are designed as 2in/2out, 3in/3out or 4in/4out configuration, but are never designed as 1in/1out. The lower the input/output port capacity, the higher is the traffic congestion (or reduced throughput capability) of the system. Example intrabay input/output ports are shown in Figure 33.15 (in the preceding section on interbay systems).
- *Production equipment load ports for FOUPs*. It is at these ports where the intrabay vehicle delivers unprocessed FOUPs for processing, and picks up finished material. In 200 mm, although there were mechanical standards for load ports, it was rarely implemented correctly. In 300 mm however, SEMI standard E15.1 load ports have become the mandatory load port configuration. Standardization of load ports is one of the main reasons why intrabay AMHS has become more pervasive and has contributed significantly to reduce the cost of ownership of the fab wide intrabay systems. A summary of 300 mm equipment interface standards is presented in Figure 33.21.
- As it can be seen, the production equipment load port is the critical physical interface point at which the AMHS and the production equipment exchanges the FOUP wafer carrier.

FIGURE 33.19 Example overhead hoist transport (OHT) vehicle (OHV). (Photograph courtesy of Brooks Automation.)

Non-proprietary, industry standards have been fundamental to the success of the 300 mm wafer fab automation. An example 300 mm tool load port is shown in Figure 33.22.

- *Lithography equipment RSP load ports for reticles*. Due to relaxation of cleanroom cleanliness as a result of tool minienvironments, this capability has seen increased use in 300 mm applications whether reticle deliveries are made by AMHS or manual means. Intrabay OHT vehicles can be configured to deliver and pick up RSPs at the lithography expose equipment. An example RSP load port with pod mechanical positioning lead-in fixtures and RF ID reader unit is shown in Figure 33.23. Typical FOUP and RSP load port configurations for a linked photoresist processor and lithography patterning tool are shown in Figure 33.24.
- *Overhead buffer (OHB)*. These carrier shelves are supported from the OHT track structures and located adjacent to the track where OHT vehicles place carriers for temporary buffering in route to their intended destinations. The OHB provides carrier storage locations without the use of cleanroom floor space and finds service as an additional carrier queue position(s) for high throughput production tools.
- *Intrabay control system*. This control system synchronizes, sequences and dispatches material movement between different production equipment in the bay and also initiates the starting of wafer processing for each equipment in the bay.
- *Front opening unified pod and RSP carrier identification systems*. These systems are responsible for reading the identification of different carriers as it passes through different elements of the intrabay material handling system. Typically, there are carrier ID readers at the input and output

FIGURE 33.20 Example overhead hoist transport (OHT) vehicle. (Photograph courtesy of Daifuku Co., Ltd.)

ports of stockers as well as each load port in the factory. The purpose of the identification system is to first read and then verify that the carrier ID on each specific load port is exactly where the host computer system thinks it ought to be. If the ID that is read is different from the information in the host system, some kind of anomaly handling is initiated that requires human intervention to resolve the data difference between the two systems.

- *User interfaces*. These are computer screens and user data input/output units located in different points in the factory, where production personnel can access or enter information on the status of wafer lots as well as get the status of the different components of the material handling system, including the intrabay vehicles.
- *Manual carts/personnel guided vehicles (PGVs)*. serve as start-up intrabay transport systems while the process and equipment are under development. Additionally they could become back-up transport systems whenever the intrabay system is down, either for scheduled maintenance or repair. The use of PGVs in an automated 300 mm factory has been shown to reduce the throughput of the line and is generally used as a last resort. In factories, where very high reliability has been achieved for the intrabay transport system, PGVs are generally not used at any time in the factory.
- *Powered conveyor systems*. These conveyor systems are sometimes used as floor based intrabay systems. When installed on the floor they are installed and run at the production equipment

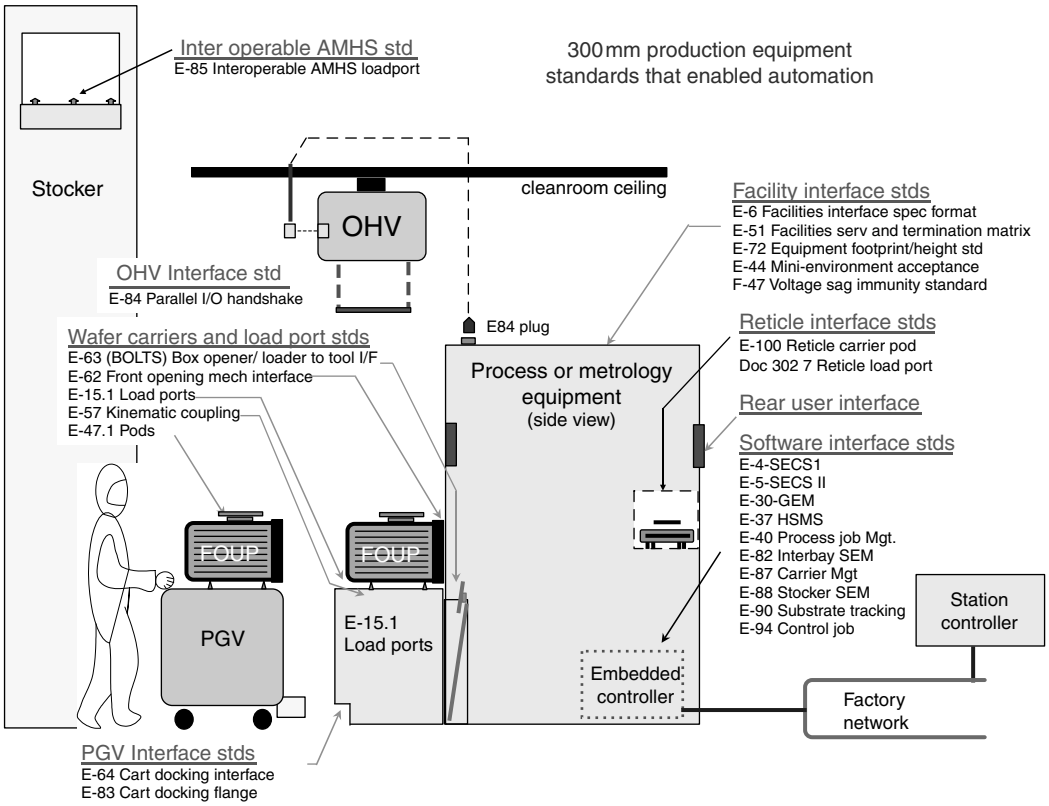


FIGURE 33.21 300 mm tool interface standards.

load port level, and rely on pick and place type robotic units to transfer carriers from the conveyor to the load port and vice versa. At least one robotic pick and place mechanism will be required for each production tool in this type of intrabay implementation. Conveyor based intrabay systems have a significant higher throughput than that of vehicle based systems. Use of conveyor system is expected to increase over time, as more and more fabs become both high volume and high mix type production environments and the factories use fewer and fewer wafers in the carrier (which results in significant higher throughput requirements).

- *Ceiling based powered conveyor systems*. These systems can also be hung from the fab ceiling. In such an implementation, vertical elevators or OHT vehicles are required for bringing the carriers down to the equipment load ports. When vertical elevators are used, there will be a need for at least one elevator for each production tool.

Figure 33.25 shows an OHV on a monorail track that is installed in the ceiling and picking up a FOUP from a tool load port. Overhead hoist vehicles are currently the most popular transport vehicle for transporting carriers between production equipment in 300 mm fabs. The unified OHT track network and its ability to route vehicles directly tool-to-tool are illustrated in Figure 33.26. Examples of other intrabay AMHS technologies are shown in Figure 33.27 and Figure 33.28. These technologies are generally dependent on interbay system transport for tool-to-tool carrier deliveries across the fab. This is because the intrabay loops span only certain regions of the factory (mostly dedicated to a single bay or a group of adjacent bays) and rely on the stocker and interbay segments to move carriers across any two points in the factory.

FIGURE 33.22 Example 300 mm process tool load port with front opening interface mechanical standard (FIMS) port and status indicator lights shown. (Photograph courtesy of Asyst Technologies, Inc.)

33.4.4 Implementation Barriers and Guidelines for Overcoming Them

Barriers to be considered when planning intrabay implementations are shown in Table 33.17. It is critically important for the AMHS design engineers to ensure all barriers are eliminated prior to automation system start-up.

Often times, designers and users of intrabay system are confronted with a choice of vehicle types as well as variations in load handling capabilities. Each type has inherent advantages and some disadvantages and must be considered before selection. Table 33.18 and Table 3.19 show key items that must be considered prior to selection.

33.4.5 Key Performance Factors for Intrabay Systems

Intrabay systems are judged by their ability to deliver wafer carriers and lithography reticles to their intended equipment load port destinations reliability 7 days a week, 24 h a day, while consistently meeting or exceeding the expected delivery time window of fab manufacturing. In addition to key

FIGURE 33.23 Example RSP load port with pod lead-in fixtures and radio frequency identification (RF ID) tag reader unit at right side. (Photograph courtesy of Brooks Automation, Inc.)

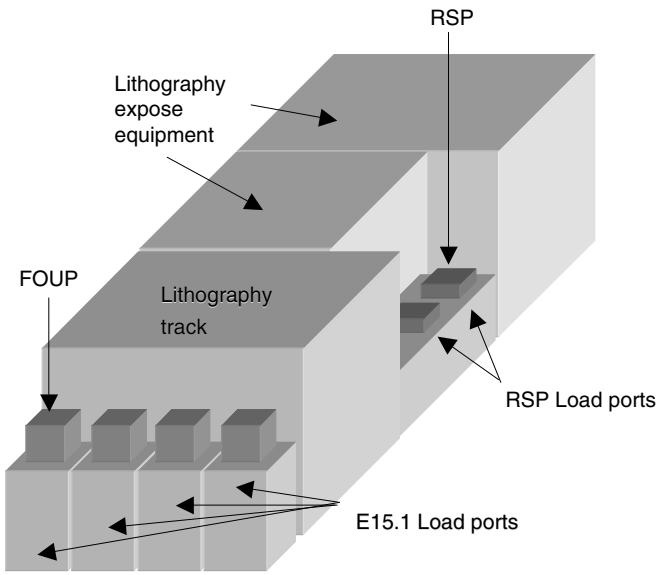


FIGURE 33.24 Typical FOUP and RSP load port configurations for an advanced lithography expose tool in a 300 mm manufacturing environment.

FIGURE 33.25 Overhead hoist vehicle (OHV) on monorail track installed in the ceiling and picking up a FOUP from a tool load port. (Photograph courtesy of Intel Corporation.)

items listed in Table 33.8 (in the interbay performance factors and their relative importance, which also apply to intrabay systems), Table 33.20 shows additional performance factors that are also very critical.

Intrabay systems must be designed to meet the throughput demands of the IC manufacturing facility for many years. Since processing changes and manufacturing requirements change dramatically at each technology node, which is driven by the pace of Moore's Law, it is critical that a holistic view of requirements be assessed prior to any intrabay solutions selection. The task is complex because the intrabay solution chosen would be required to operate for 5–10 years, despite that fact the factory

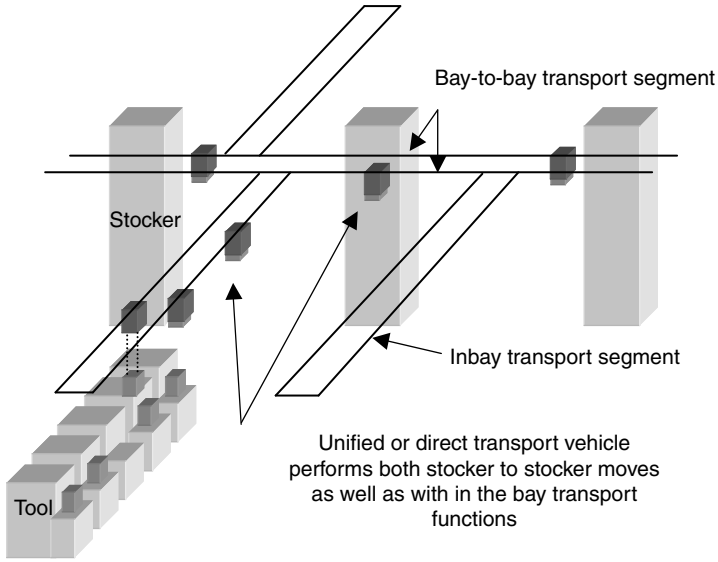
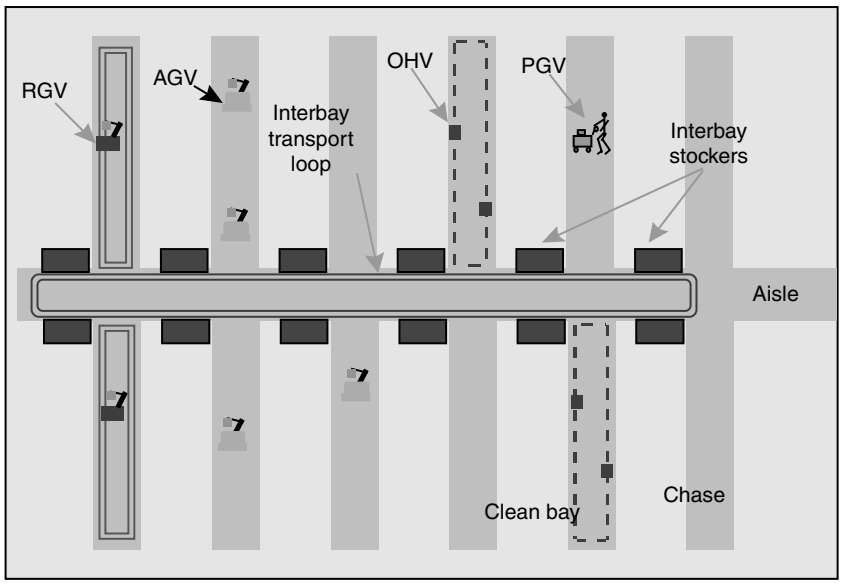

Bay-to-bay transport segment
Stocker
Inbay transport segment
Unified or direct transport vehicle performs both stocker to stocker moves as well as with in the bay transport functions
Tool

FIGURE 33.26 Unified transport track system can be used to transport carriers between stockers in the factory as well as to and from any production tool in the factory layout.



AGV = Automatic Guided Vehicle (200 mm & 300 mm)
RGV = Rail Guided Vehicle (200 mm & 300 mm)
OHV = Over head Hoist Vehicle (300mm)
PGV = PersonGuided Vehicle or Cart (200 mm & 300 mm)

FIGURE 33.27 Different intrabay technologies and their interface with interbay systems. Generally only one type of intrabay solution is chosen for each bay to eliminate transport system interferences.

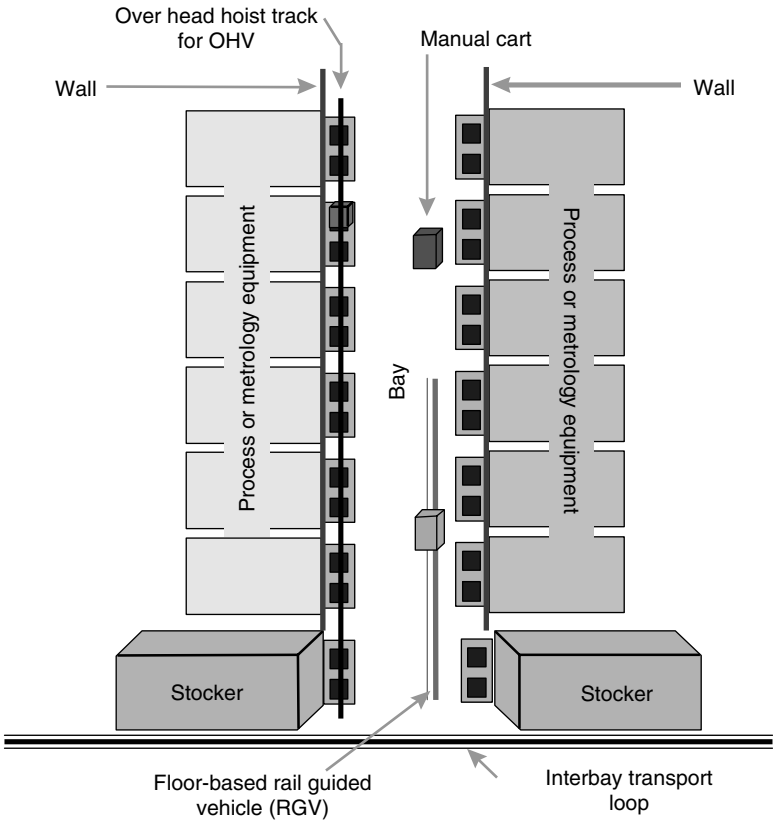


FIGURE 33.28 Close-up view of different intrabay technologies in relation to interbay automated material handling systems (AMHS).

undergoes dramatic changes in equipment models, operational requirements and manufacturing complexities, including material routing requirements, every 2–3 years.

In such a dynamically changing manufacturing environment, several critical requirements drive the operational attributes of the intrabay system. Frequent production equipment change outs every few years mandate that the intrabay system be flexible enough to handle these variations. With advent of widespread use of 300 mm mechanical interface standards, equipment changes are felt must less on the transportation system now than during the 200 mm era. However the equipment load port placement axes is critical for an existing transport system to be able to successfully load and unload it. It is for that reason that AMHS vehicle suppliers provide a broad set of solutions that can help alleviate some of the complexities if the equipment install accuracy is lower. Multi-axis robots on vehicle with multiple degrees of freedom enable one to overcome some of the above problems, but it is at the expense of higher cost and lowered cycle times. Another critical component is throughput capability. Due to the changing needs over time, an existing production tool could easily get replaced with a new tool that has a much higher throughput. This puts higher throughput demands on the transport system serving that bay. One way to successfully counteract this threat is to ensure that the transport system has the highest throughput possible, at the time of the selection. Other parameters that are equally important are AMHS space efficiency (smallest possible footprint desired), handling and routing flexibility for the transport vehicles and tracks, meeting all safety requirements and operating well within the cleanliness requirements as defined by the process and the factory.

TABLE 33.17 Common Barriers to Successful Intrabay Automation and Some Possible Remedies

Barriers to Intrabay	Possible Pitfalls and How to Avoid Them
The load port of production tools not installed in a straight line along a bay side	It is important for the center line of the E15 load ports to be installed along a straight line within a ± 10 mm accuracy. This will permit the front opening unified pod (FOUPs) to be sitting on the production equipment load port within the capture range of the overhead hoist transport (OHT) vehicle pick-up and set-down arms
The production equipment is not uni-carrier (i.e., the equipment returns wafers into another FOUP and not the original FOUP)	Attempting intrabay on production equipment that do not have uni-carrier capability is extremely risky and most likely prone to failure. This is because the intrabay vehicle has to perform the FOUP swapping (i.e., return the empty FOUP from one port to the other), and since it has to maintain empty and full carrier tracking, the control system and its associated anomaly handling become very complex and unwieldy. The only exception to the uni-carrier rule is wafer sorter equipment where wafers are physically sorted and placed in separate carriers intentionally
The production equipment requires significant number of send-ahead wafers or test wafers that interrupt normal production	Intrabay automated material handling systems (AMHS) will be very difficult to implement and operate efficiently in such an operating model. Efforts to scale back the use of send-ahead wafers and manual transport of these wafers should be considered. Often times, in such a situation of excessive send-ahead wafers, the intrabay transport system will get easily overwhelmed with additional moves to be performed, that the delivery time of other material will get negatively impacted
The production equipment requires excessive front access for tool PM and/or repair	Intrabay AMHS will be very difficult to operate effectively. Intrabay should not be attempted in such an environment
The production equipment or process is characterized by extensive downtime and requires frequent operator assists	Intrabay AMHS will be very difficult to operate effectively. Intrabay should not be attempted in such an environment
No buffering—i.e., the production equipment does not have additional load ports or internal carrier buffer locations	Typically, intrabay automation should be attempted only on those equipment which have an additional load port or internal buffers. Otherwise, the production equipment is likely to wait (remain idle) for the intrabay vehicle to deliver its next carrier. Equipment buffers must be sized to handle an average delivery time of 5 min or a peak (95% of all lots will be delivered) in 18 min. See the delivery time frequency graphs in Figure 33.29

The intrabay technology chosen which includes the vehicle type, the load handling features and the transport loop type all play a very important role in this. Some of the important specifications for the most commonly used intrabay vehicle systems are shown in Table 33.21 and Table 33.22.

Table 33.22 presents metrics for overhead intrabay delivery equipment. This equipment is generally qualified to Class 10 at 0.2 μm or better airborne particle cleanliness.

33.4.6 Intrabay System Throughput Constraints and Possible Solutions

In fabs with interbay and intrabay AMHS, stocker input/output ports interface with the interbay and intrabay transport vehicles. The intrabay vehicles in turn interface with production equipment during the actions of loading and unloading FOUPs at the production equipment load ports. With this integration challenge, it is important to know and understand which elements of the overall system are likely to pose throughput "bottlenecks" for the entire system. Table 33.23 summarizes the problem areas and suggests methods for reducing these constraints.

TABLE 33.18 Practical Implementation Guidelines and Relative Comparisons of Intrabay System Configurations

Vehicle Type	System Throughput	Time to Position for Load and Unload	Routing Flexibility	Notes and Comments
Rail guided vehicle (RGV) on the floor (some suppliers have a continuous loop rail, others have straight rail)	Very good	Very good	Less flexible	Those suppliers that have a loop rail can put more than one vehicle on the track, further increasing system throughput. Rail cost has to be added to vehicle(s) cost
Tape guided automated guided vehicle (AGV)	Good	Good	Very good	To further increase system throughput, more than one vehicle is used in a bay. No need for rail costs, but battery charging system may have to be included
Free roving AGV	Good	Poor	Very good	To further increase system throughput, more than one vehicle is used in a bay. No need for rail costs, but battery charging system may have to be included
Overhead hoist vehicle (OHV)	Very good	Good	Less flexible	More than one vehicle can be installed on the loop to increase bay throughput. Rail cost has to be added to vehicle cost
Powered roller conveyor	Highest	Very good	Less flexible	Carrier load/unload mechanism(s) are required at each load port of equipment

33.4.7 Intrabay System Sizing and Planning

Design and evaluation of intrabay system capability must incorporate the manufacturing process needs and throughput requirements of the different tools in a particular bay. They are different depending upon the layout, the individual equipment throughput requirements, the number of carriers batched and run, and the material routing rules used by the transport system.

The planning and intrabay technology selection procedure consists of the following steps. Each is accompanied by a numerical example for clarity.

1. Determination of peak sustained throughput of the intrabay system
2. Completion of the production equipment interface audit to develop a list of feasible technology options
3. Developing feasible solutions combining the throughput requirements and list of feasible intrabay technology options

TABLE 33.19 Practical Implementation Hints—Relative Comparisons of Robot Types Used on Intrabay Vehicles

Carrier Handling Mechanism	Cycle Time to Load or Unload	Placement Versatility[a]	Comments
Multi-axis robot	Slower	Excellent	Highest cost
Simple 2-axis shuttle arm	Very fast	Limited	Lowest cost
Hoist suspended end effector	Fast	Good	Medium cost
Pick and place shuttle arm that lifts carriers from powered roller conveyors	Fast	Limited	Lowest per unit cost, but highest total system cost due to the need for units at each production equipment

[a] Versatility is defined as the capability of the robot arm to load and unload a wide variety of load port configurations and when the load placement orientations are significantly different.

TABLE 33.20 Intrabay Performance Factors and Their Importance

Performance Factors	Why Is It Important?
Delivery time -stocker to equipment load port	This is defined as the time required for the intrabay system to transport a carrier from a stocker shelf to the load port of the requesting (destination) equipment. In addition to the throughput capability of the transport system, other equally important drivers of this metric include the transportation distance the carrier has to move along the transport path, as well as processing batch size of the destination equipment. If the intrabay system is not sized adequately, longer than expected delivery times can significantly (negatively) impact the throughput of the intrabay system, resulting in reduced production throughput of the line. This effect can have crippling impact to factory output as factory constraint equipment is serviced by this transportation system
Delivery time -equipment load port to stocker	This is defined as the time required for the intrabay system to transport a carrier from the load port of production equipment to a stocker that is connected to the same intrabay transportation system. The critical elements that drive this delivery time performance are similar to the ones described above. The goal is to have the shortest delivery time while meeting the peak move rate required of the intrabay system
Load port exchange time (LPET)	This is defined as the time elapsed from when a carrier is requested to be retrieved from a specific load port of a production equipment until the time when another carrier is successfully placed on that same load port. Among all metrics for intrabay systems, this performance metric is unquestionably the most critical metric for factory performance. The longer the LPET duration is, the greater is the probability for production equipment to be "starved" of production material, resulting in reduced output from that equipment. Short LPET durations are extremely vital for those production equipment that have very short processing durations, such as metrology equipment.

TABLE 33.21 Comparison of Floor Based Intrabay Vehicle System Metrics

Types	Carrying Capacity	Cleanliness Class	Typical Transport Speed Ranges (m/min)	Positioning Time Prior to Load or Unloading (s)	Turning Radius
Free roving automated guided vehicle (AGV) with multi-axis robot	2 Front opening unified pod (FOUPs)	Class 1 @ $\geq$ 0.1 μm	18–60	5–17	Some $\geq 2\times$ vehicle length; others can spin turn
Tape guided AGV with multi-axis robot	2 FOUPs	Class 1 @ $\geq$ 0.1 μm	18–60	5–12	Some $\geq 2\times$ vehicle length; others can spin turn
Straight rail guided vehicle (RGV) with multi-axis robot	2 FOUPs	Class 1 @ $\geq$ 0.1 μm	30–60	1–3	Turning not possible
Loop RGV with multi-axis robot	2 FOUPs	Class 1 @ $\geq$ 0.1 μm	30–60	1–3	$\geq$ 0.5 m
Straight RGV with simple three axis shuttle arm	2 FOUPs	Class 1 @ $\geq$ 0.1 μm	30–60	1–3	Turning not possible

TABLE 33.22 Intrabay Transport Systems (Overhead Mounted)

Types	Payload Capacity	Hoisting Robotics DOF; Vertical Stroke; Transfer Cycle Time	Typical Transport Speed Range	Typical Track Turning Radii	Typical Track Routing Mechanisms
Overhead hoist transport (OHT) track and vehicle (OHV)	1 carrier, 10–15 kg per vehicle	4 degrees of freedom; 2.6–2.8 m typical; 12–15 s	30–180 m/min	Monorail with 0.3–0.5 m minimum	Merge/diverge track; Turntable
Powered roller conveyor	Sequential queuing of carriers using accumulation=/> 10 kg/carrier	Carrier elevators/lifters located at tool load ports; 4 degrees of freedom; stroke as required; 18–24 s	12–18 m/min adjustable	Right angle turning possible	Powered rollers with carrier travel directions at 90° to each other

TABLE 33.23 Automated Material Handling System (AMHS) Equipment Interfaces and Constraint Areas

Equipment Elements	Interfaces to	Typical System Throughout Limiters or Constraints	Practical Methods for Increasing Throughput
Stockers	Overhead monorail track	Overhead track transfer cycle time	Use additional transfer units, or multiple positions at a stocker for front opening unified pod (FOUP) exchange
	Automated guided vehicles (AGV)	Input/output port capacity	More ports and/or queues (2in/2out, 3in/3out or 4in/4out)
	Operator personnel	Input/output port capacity	More ports and/or queues (2in/2out, 3in/3out or 4in/4out)
	Overhead hoist vehicles (OHV)	Overhead track transfer cycle time	Use 2in/2out ports (as a minimum) or multiple positions at a stocker for FOUP exchange
	Rail guided vehicles (RGV)	Input/output port capacity	More ports and/or queues (2in/2out, 3in/3out or 4in/4out)
AGV	Process, metrology, and other AMHS equipment	AGV parking and battery charging Vehicle docking cycle time Load/unload cycle time	Use opportunity charging (vehicle gets charged when idle) Add additional vehicles to eliminate bottleneck
RGV	Process, metrology, and other AMHS equipment	If straight rail, then only one vehicle can be used	Use loop rail with multiple vehicles on the loop
OHV	Process, metrology, and other AMHS equipment	Vehicle load/unload cycle time Only one carrier per vehicle	Use track siding to reduce congestion or multiple positions at a stocker for FOUP exchange Add additional vehicles E15 load port compliance
Inter-level lift	Overhead monorail track Stocker I/O port conveyors overhead monorail track	Inter-level lift cycle time Single carrier lift load carrying capacity	Shorter cycle time lift, additional lift or transfer unit Multi-carrier capacity lift

33.4.7.1 Calculation of the Peak Sustained Throughput of the Intrabay System

Assuming the fab operates non-stop 168 h per week, the average throughput T of the intrabay system is computed by the following formula:

$$T = W(N_1 + N_2 + N_3)/(L \times 168) \text{ carrier moves/h}$$

where

- W fab wafer starts per week,
- L transport lot size (in wafers),
- N_1 number of times in the process flow a lot moves from the stocker to a production equipment in that loop,
- N_2 number of times in the process flow a lot moves from one production equipment in that intrabay loop to another equipment in the same loop, and
- N_3 number of times in the process flow the lot moves from the production equipment back to the stocker, all located on the same loop.

As an example, if $W = 5000$ wafer starts per week, $L = 25$ wafers, $N_1 = 10$ times, $N_2 = 9$ times, and $N_3 = 10$ times, then, $T = 5000\ (10 + 9 + 10)/(25 \times 168) = 34.5$ carrier moves/h. Note: This number is not to be used in the intrabay design.

Typically, the peak throughput of a bay is generally 1.5—2.0 × the average throughput.

If you design to a peaking factor of 1.5, then the peak throughput requirements $= 1.5 \times 34.5 = \sim 52$ carriers/h.

If you design to a peaking factor of 2.0, then the peak throughput requirements $= 2 \times 34.5 = 69$ carriers/h.

33.4.7.2 Production Equipment Interface Audit

Table 33.24 summarizes key equipment interface parameters that have to be verified on each production equipment prior to intrabay implementations. The table also shows what drives the use of different intrabay technologies in 300 mm as well as 200 mm wafer processing fabs.

33.4.7.3 Combining Technology Feasibility and Throughput Requirements to Arrive at Possible Options

From Table 33.24, it can be seen that there are several different equipment features that affect the feasibility of selecting a particular intrabay technology. It can also be inferred that if there are several different types of equipment in a bay, then there is high likelihood of the attributes being different. These need to be considered upfront when designing and implementing intrabay systems. If there are only a few different equipment types in a bay, or it is predominantly one type of equipment or if the equipment load ports and interfaces meet the SEMI standards, such as E15, then it is possible to select more than one option as an intrabay vehicle candidate. Until compliance with SEMI standards is achieved, the ability to pervasively implement intrabay technologies is limited. For 300 mm, there is widespread equipment configuration compliance to meet intrabay functionality. Consequently, most 300 mm fabs have pervasive and fab-wide intrabay automation operational. This success was achieved by having interface standards and measurement criteria for ensuring standards compliance early in the development cycles of the production and automation equipment. As a result, the OHT vehicle (with its very limited degrees of freedom) is able to flawlessly support the load and unload requirements of each and every production tool in a modern 300 mm fab today. This is the main reason why the cost effectiveness of 300 mm intrabay is easier to demonstrate as long as the implementation risk is minimized.

The ability of OHT track to be unified or "networked" across all the tools on a given level of the fab gives it the important capability to support high carrier move rates, lower carrier delivery times, and reduce variability in delivery times. This OHT capability of rapid direct tool-to-tool delivery is illustrated

TABLE 33.24 Compatibility of Production Equipment and Intrabay Vehicle Attributes

		Intrabay Vehicle Compatibility with the Particular Attribute					
No	Critical Production Equipment Attributes that Impact Intrabay Success	Automated Guided Vehicle (AGV) with 5 or 6-Axis Robot	AGV with Simple Shuttle	Rail Guided Vehicles (RGV) with 5 or 6-Axis Robot	RGV with Simple Shuttle	Overhead Hoist Vehicle (OHV)	Comments
1	Front face of load ports cannot be lined up in a straight line	Yes	Yes	No	No	No	Not a issue in 300 mm
2	Load ports of all tools are not at the same horizontal level	Yes	No	Yes	No	Yes/No	Not a issue in 300 mm
3	Tools require front maintenance near the load ports	No	No	No	No	Yes	Not an issue in 300 mm
4	Front face of equipment is not flush with the bay wall	Yes	Yes	Yes	Yes	No	OHV not feasible. Not an issue in 300 mm
5	Equipment does not have SEMI E23 Parallel I/O interface (for ground based vehicle and SEMI E84 for overhead hoist transport (OHT) vehicles)	No	No	No	No	No	Intrabay not feasible. Not an issue in 300 mm
6	Equipment cannot be remotely controlled by a cell or station controller	No	No	No	No	No	Intrabay not feasible
7	There is no room for vehicle and/or rail run-outs on either ends of the bay	Yes	Yes	No	No	No	RGV, OHV rails need run outs

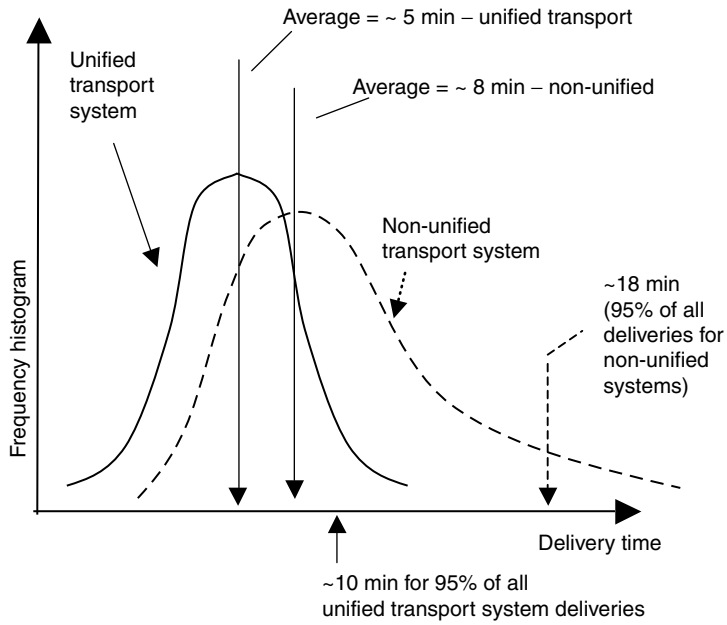

Average = ~ 5 min – unified transport
Unified transport system
Average = ~ 8 min – non-unified
Non-unified transport system
~18 min (95% of all deliveries for non-unified systems)
Frequency histogram
Delivery time
~10 min for 95% of all unified transport system deliveries

FIGURE 33.29 Graph showing example distribution of intrabay delivery times. Non-unified transport systems are generally slower than unified transport systems from a delivery time perspective.

in Figure 33.29 in comparison with non-unified transport systems, which are slower and operate with higher variability in carrier delivery times.

33.4.8 300 mm Process and Support Equipment Interfaces

Standardized mechanical, electrical, and communication interfaces along with physical easements for loading/unloading wafer carrier at production equipment load ports are key "enablers" for cost effective intrabay AMHS. Wafer logistics interfaces and easements were first outlined and issued as guidelines for 300 mm production equipment by the International 300 mm Initiative (I300I) [3,4] and also in conjunction with the Japan 300 mm Working Groups (J300) consortium [5]. The SEMI standards related to 300 mm tool interfaces with AMHS, factory control systems, and wafer fab facilities are summarized in Figure 33.21. Compliance to SEMI Standards is used by IC device manufacturers and their suppliers for evaluating 300 mm process and metrology equipment readiness to support cost effective intrabay implementations [7].

Status indicator lights on the 300 mm tool load port can be used like a traffic light to coordinate operator access to the load port with intrabay AMHS FOUP pick-up and delivery operations. Common user implementations and functionalities of these lights are outlined in SEMI Auxiliary Document AUX-006. An example implementation of the status indicator lights is shown in Figure 33.30.

Integrated circuit manufacturers and equipment suppliers have both realized that cost effective and timely standardization activities in the equipment interface area can greatly reduce 300 mm equipment development costs and minimize the number of different options that have to be developed and tested, prior to factory implementation. Through the help of SEMI, many interface standards have been developed. These standards when implemented on production equipment (process and metrology equipment), have greatly simplified intrabay functionality and implementation.

Table 33.25 shows the most critical interfaces, standards and their impact to intrabay automation, along with the typical pass/fail criteria to ensure compliance to SEMI standards:

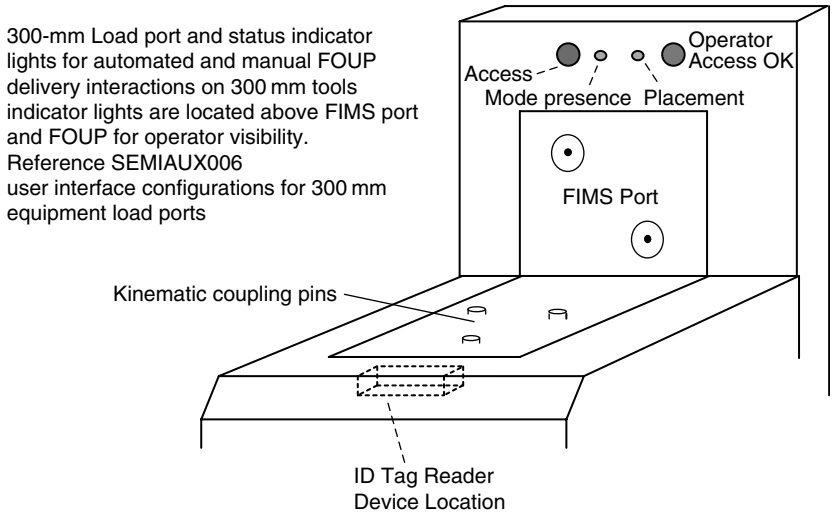

300-mm Load port and status indicator lights for automated and manual FOUP delivery interactions on 300 mm tools indicator lights are located above FIMS port and FOUP for operator visibility. Reference SEMIAUX006 user interface configurations for 300 mm equipment load ports
Operator Access OK
Access
Mode presence
Placement
FIMS Port
Kinematic coupling pins
ID Tag Reader Device Location

FIGURE 33.30 Example status indicator lights on a front opening unified pod (FOUP) load port.

33.4.9 Buffering on the Production Equipment for Uninterrupted Processing:

A very critical component for uninterrupted processing on the production tool is to ensure adequate FOUP buffering locations exist on the tool in order to ensure the output of the tool is not negatively impacted by the AMHS delivery time variations. During normal operations, and while the tool is processing wafers, the intrabay vehicle can opportunistically transport the FOUP(s) that is to be processed next in sequence to the buffer location, before that FOUP is needed for processing. In such an operating model, the production tool is able to run in an uninterrupted manner, enabling continuous processing and highest equipment utilization.

A simple equation can be used for estimating the number of additional locations or buffer capacity on the tool where FOUPs can wait for processing. The buffer quantity is dependent on the peak delivery time of the AMHS transport vehicle, the production runrate of the tool, and the typical processing batch size of the tool. A simple equation helps one determine the number of additional buffering locations on a production or metrology tool:

$$\text{Additional buffer(carrier) locations} = \frac{\text{Peak delivery time (in minutes)} \times \text{Tool runrate (in wafers per hour)}}{\text{Processing batch size (in wafers)} \times 60 \text{ min\h}}$$

Example: if Peak delivery time of the transport system is 20 min, tool runrate is 70 wafers per hour, and processing batch size of the tool is 25 wafers, then

$$\text{Additional buffer (carrier) locations needed} = \frac{20 \times 70}{25 \times 60} = 1 \text{ carrier (rounded up)}$$

33.4.10 Benefits of Intrabay Systems

The benefits of intrabay systems come from several factors which are summarized in Table 33.26.

TABLE 33.25 Automated Material Handling Systems (AMHS), and Integration Interface Tests, Related Standards Pass/Fail Criteria

#	Interface Test and Applicable SEMI Standard	Test Description	Key Items to Check for Pass/Fail Criteria
1	Front opening unified pod (FOUP) compatibility: (SEMI E1.9, E57–E15.1)	Ensure the load port(s) of the production equipment inter-operates with any SEMI E47.1 compliant FOUP	Correct registration of FOUP with kinematic coupling pins on load port Verification of all clearance dimensions in front and on sides of the FOUP when placed on load port FOUP placement and presence sensing and individual enunciation
2	SEMI E15.1 compliance testing of load port	Ensure load port compliance with all dimensions and access rules	Conformance to all critical dimensions: S, C1, C2, H, H1, H2, D, and D1 Carrier fork lift exclusion zone test Load port orientation tests FOUP hold-down test Front opening interface mechanical standard (FIMS) interface verification FOUP door opening and closing
3	Automation delivery easement verification (SEMI E15.1)	To verify the production equipment has clearances that permit an overhead hoist vehicle (OHV) to load or unload a FOUP at the load port	Load port meets D, D1, H2, and S critical dimensions No obstructions in the "chimney" zone above the load port for OHV operation No obstructions in front of the load port for rail guided vehicles (RGV)/automated guided vehicle (AGV) or personnel guided vehicle (PGV) operation
4	Cart docking interface exclusion zone (SEMI E64)	To verify the load port and the front face of the production equipment has the clearance for manual cart (or PGV) docking	Exclusion zones Lh and Ld exist at load port areas Exclusion zones extend along the full width of the equipment
5	Equipment buffering of carriers	To verify that the production equipment has met the minimum carrier buffering criteria as defined by I300I guidelines	Vertical diffusion furnaces: 12–16 carriers plus 2 E15.1 load ports Auto wet stations: 10–12 carriers plus 2 E15.1 load ports Linked litho equipment: 4 E15.1 load ports Ion implanter: 4 E15.1 load ports All other process equipment: 2 E15.1 load ports All in-line metrology equipment: 2 E15.1 load ports Low use (off-line) metrology equipment: 1 E15.1 load port
6	Straight line installation dimensional verification	To verify that the equipment is capable of being installed in a straight line	Critical dimensions of the equipment as measured from the Equipment Datum point must meet the customer install tolerance This is specific for each type of equipment
7	Light stacks	To verify that the equipment has two status light stacks (or signal towers), one located in front and one in the rear	Operational verification of light colors signifying specific equipment conditions To check the option whether the accompanying buzzer sound can be deactivated at user request Tops verify that the light stack does not protrude into the OHV easement zone or into the AGV/RGV/PGV travel paths

(*continued*)

TABLE 33.25 ***(Continued)***

#	Interface Test and Applicable SEMI Standard	Test Description	Key Items to Check for Pass/Fail Criteria
8	Alternate equipment user interface location	To document locations of the primary and the secondary user interfaces of the equipment, and that only one is functional at any given time	Verify the presence and functionality of the primary user interface Verify the presence and functionality of the secondary user interface Verify only one interface is operational at any given time Verify the front user interface location and access does not conflict with OHV clearances and RGV/AGV/PGV travel paths
9	Carrier and slot Integrity	Ensure that the production equipment is capable of returning wafers after processing to the same carrier and into the same slot from which they originated	Cycle the equipment so that all wafers after processing return to the original carrier and into the same slot it originated from prior to processing
10	SEMI E84 parallel I/O photo-coupled interface functionality and clearances. (For ground based systems, the standard is E23)	Ensure functionality of the E84 photo-coupled interface for low-level communication handshake to the intrabay transport vehicle	Ensure that the interface uses the photo-coupled option and not the wire option Ensure E84 operational functionality Ensure an interface exists at the specified location on each load port for interface to either an RGV or an AGV (using E23) For interface to an OHV, ensure the capability exists for installing this interface at a height of about 3.4 m above the floor for each load port

TABLE 33.26 Intrabay Systems Benefits and Methods of Demonstration

Benefit Category	Demonstration Method	Rationale and Comments
Improve operational rate of key production equipment by minimizing the dead-time between sequential lot processing	Intrabay area consistently outperforms the manual area	The intrabay system is capable of keeping the production equipment always running without interruption by utilizing pre-process queues (buffering). 3%–10% increased output and cycle times may be achieved and have been reported on key constraint equipment
Direct labor savings, and reduction in human induced variability	Number of direct labor in the intrabay area vs. the number of people in the manual area	Direct labor savings between 25%–50% may be possible. This savings is restricted to the number of people previously utilized in manual work-in-progress (WIP) movement tasks. No decrease in equipment maintenance functions can be attributable to intrabay operations
Reduction of operational mistakes—improved line and die yield	Fewer wafers scrapped	Consistent way of running the production areas without respect to the time of day or shift. The correct lot is always dispatched to the correct tool each time

33.5 Material Control System

33.5.1 Purpose and Typical Applications

The MCS initiates and coordinates concurrent movement of wafer carriers (and/or reticle carriers) within the AMHS. It is also responsible for coordinating interbay and intrabay system activities between stockers and production equipment load ports. The MCS maintains real time status and monitoring of interbay and intrabay AMHS components, and also maintains an accurate inventory database of wafer carrier locations within the AMHS. Identification systems, responsible for reading carrier ID at key entry and exit points within the AMHS, also interface with the MCS to provide carrier identity and location information.

The MCS accomplishes its functions by:

- Interfacing with the fab MES from where it receives wafer lot destinations and priority information and it updates the MES with lot inventory locations after the lot movements are successfully completed.
- Interfacing with the various elements of the AMHS where it receives carrier and lot ID upon carrier entry to the AMHS and coordinates and commands the movement of lots by the AMHS mechanisms to their destinations as designated by the dispatch rules and priorities of the MES.
- Providing reports on lot movements and locations, lot inventory summaries, non-product inventories, and data on AMHS reliability.

Material control system is therefore a part of an overall, mission critical, manufacturing control and execution system that manages and carries out the production material logistics in the order defined by production control priorities of the factory. Figure 33.31 shows a very simplified sketch of some of the critical software systems capabilities that exist in a modern wafer fabrication facility that manufactures 300 mm wafers in high volume.

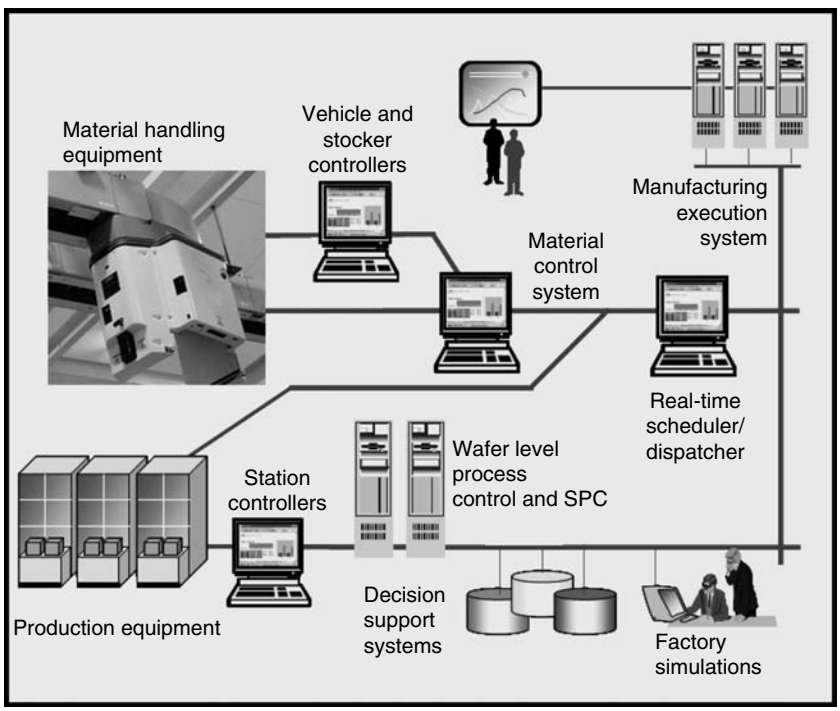


FIGURE 33.31 High level schematic of the integrated automation capabilities that manage production control, scheduling, dispatching and process control functions in a leading edge integrated circuit (IC) manufacturing factory.

33.5.2 MCS Elements, Functions and Key Requirements

Key elements, functions, and requirements of MCS are summarized in Table 33.27. Each element in the MCS system shown requires very close synchronization with each other element and must be optimized for highest performance. Optimization for highest performance is needed in order for the system to execute commands at millisecond priority interrupt levels. As a result, performance characterization at peak operating loads is an important part of system acceptance testing.

TABLE 33.27 Key Material Control System (MCS) Elements, Functions and Requirements

MCS Elements	Functions	Key Requirements
MCS application	A real time software application (generally running distributed on multiple computers) responsible for managing and coordinating material movement and tracking functions of the interbay and intrabay automated material handling system (AMHS)	Very high reliability (> 99.95%) Scaleable to higher wafer starts and throughput Quick response to system and operator queries User friendly operator interfaces Fail over capability without operator intervention Reconfiguration capability without system shutdown Be able to inter-operate with different interbay and intrabay AMHS equipment
MCS hardware	Computers (typically client -server type) on which the distributed MCS application runs	Multiple hardware for fail over and back-up Redundant hardware preferred Multi-processor computer configuration High input/output (I/O) bandwidth capability
MCS operating system	The operating system on which the MCS clients and servers run on	Multi-threaded operating system Fault tolerant capability is an added plus Disk shadowing/mirroring capability Support for multi-processor hardware
Real-time database	A real time relational database structure that maintains logical and physical lot information as well as the status and reports of all interbay and intrabay components	Multi-threaded database structure Support disk shadowing/mirroring capability (using primary and backup configuration) Be able to synchronize databases efficiently during failure event High disk I/O bandwidth capability Ability to make changes to database schema easily
Application program interface (API)	This is a standardized interface used to write interface code between the MCS and other external (3rd party) systems with which it interfaces	MCS API to the host manufacturing execution system/computer-integrated manufacturing (MES/CIM) system MCS API to real time scheduling and dispatching systems MCS API to the stocker controllers MCS API to interbay transport controller MCS API to bay or cell controllers MCS API to intrabay controllers MCS API to 3rd party lot ID tracking systems
Graphical user interfaces	Easy to use operator interface through which fab users and maintenance personnel interact with the AMHS system	Logically arranged screens, which are easy to access and consistent with each other Drill down capability—start at top level and be able to go down multiple levels of detail Be able for users to easily change what they want displayed
AMHS network or local area network (LAN)	Typically, this is a dedicated LAN used just for AMHS inter-equipment and systems transaction and message passing	Very high reliability (100%) with no single points of failure Very high bandwidth (100 MB/s) Very quick response time (< 0.25 s/transaction) Peak network utilization (< 5%)

33.5.3 Typical MCS Configuration

A typical MCS configuration which includes interbay and intrabay AMHS capability is shown in Figure 33.32.

The interbay MCS synchronizes and manages material movement between different processing bays or stockers in a factory. The interbay transport function is also a part of MCS functionality. The MCS also maintains the real time correspondence between carrier ID and lot ID while the carrier is under the control of the interbay MCS. The application is also responsible for system logging and reporting, start/up and shutdown sequences, and alarm monitoring and messaging.

The intrabay MCS, through a bay or cell controller within a bay, is responsible: for scheduling and directing carrier transport functions of the intrabay vehicle; maintaining intrabay equipment control functions (start, stop, pause, collect data, etc); maintaining communications between the production equipment and the cell controller interfaces using SECS-II messages; and providing the primary user interface for the production operator in the bay. The bay or cell controller is also responsible for maintaining the correspondence between carrier ID and lot ID while the lot is in the bay undergoing processing.

A dedicated AMHS LAN is utilized in the factory for the transmission of messages between the different MCS client server computers as well as the different stockers, interbay transport controller, and the intrabay vehicle controllers. As a result, the AMHS local area network (LAN) carries only AMHS related message traffic, and is not overwhelmed with traffic from the host MES.

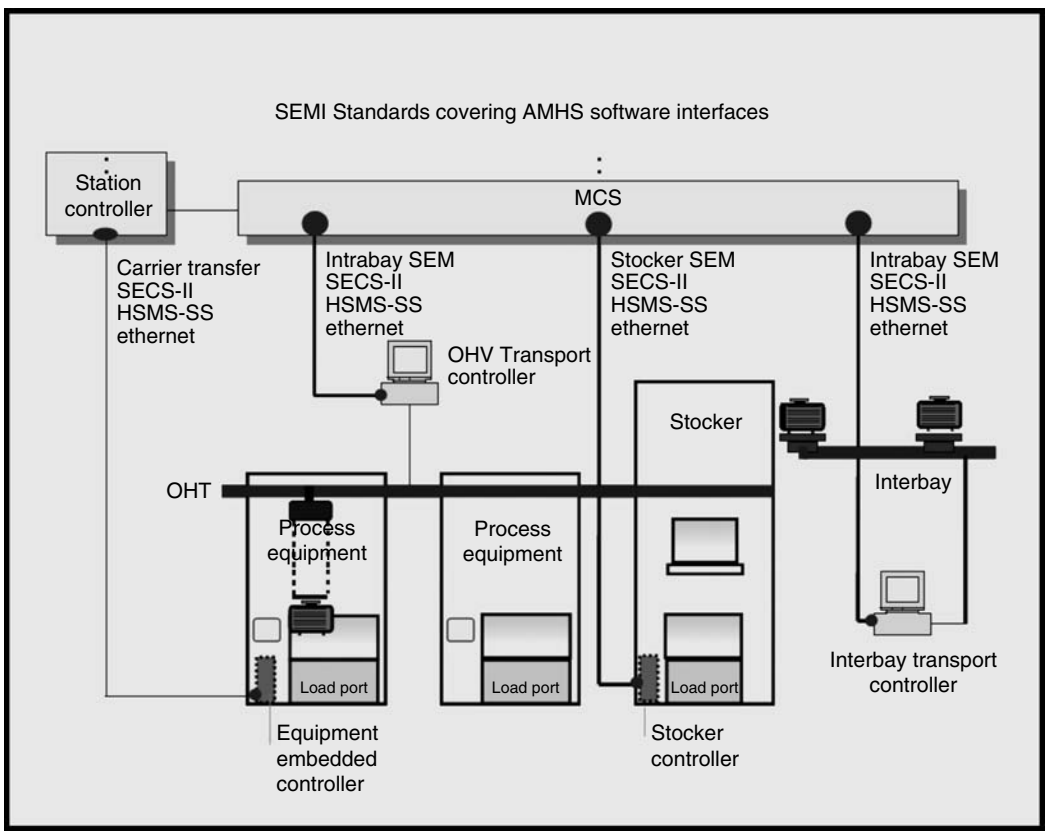


FIGURE 33.32 Typical interbay and intrabay controls configuration.

33.5.4 Key MCS Performance Factors

Key MCS performance factors and typical requirements are summarized in Table 33.28. These factors imply the need for highest levels of reliability and uptime, since the entire factory is fully dependent on the 300 mm AMHS system for material storage and transport.

TABLE 33.28 Material Control System (MCS) Performance Factors, Their Importance, and Typical Requirements

Key Performance Factor	Importance	Typical Requirement
MCS reliability	Since the MCS is the nerve center of the overall automated material handling system (AMHS) system, integrated circuit (IC) manufacturers expect very high reliability of the installed system. (This includes the combined uptime of the application and the associated computer and databases systems)	Uptime $>$99.95%
Automatic fail over	Users expect the back-up MCS computer to automatically switch on and take over the system controls in the event the primary computer(s) fails. This capability is very critical because in a very large AMHS system that moves hundreds of carriers per hour, a small unscheduled failure results in hundreds of carriers stranded and physically and logically unsynchronized with the database	Automatic failover which is a transfer to a safe state without operator intervention
MCS response time	Response times are very important because they indicate the extent of upside, the MCS has for message and transaction processing. Response time is defined as the average and peak times measured from the instant an AMHS component sends a work completion message to the time the MCS sends back an acknowledgement of the completion	Average response time $<$ 0.25 s, peak response time $<$ 1 s
MCS re-configuration without system shutdown	Such capability enables users to make configuration changes to the MCS without bringing the entire MCS down to install the software patch. This capability enables the system to run without pauses or stopping while the change is made. Configuration changes can mean adding more interbay or intrabay vehicles, bringing a new stocker on-line, or making minor changes to track routing.	No need to bring the MCS down while the changes are implemented
Real time status display and monitoring	In a large AMHS system, it is very important for system sustainers and design engineers to know the real time state of each component of the system. There is also the need to know how many and which messages are being executed and which are queued for subsequent execution. This will serve as a valuable aid for debug and troubleshooting	Real time update of all critical AMHS components. Display option based on user defined configuration
Real time inventory data base query	Be able to obtain the inventory status of carriers waiting at each stocker, at each production equipment load port, or on the interbay and intrabay transport vehicles	Data should be accessible from either interbay screens or cell controller screens
Efficient error recovery	System should be capable of rapidly recovering from either an MCS related error or AMHS equipment related error	Minimum amount of manual data entry to recover from typical system errors

33.6 AMHS Reliability and Maintainability Requirements

Device manufacturers view the interbay and intrabay AMHS to be analogous to any utility system they use in the cleanroom, such as the electric power supply or the process cooling water (PCW). They expect it to be operational all the time. Whenever such systems fail, the implications to manufacturing are severe. Consequently, AMHS systems must demonstrate very high reliability and low maintainability once they are installed in the factory. The reliability metric for transport equipment is Mean Moves between Failures, MMBF and the metric for storage equipment is mean cycles between failure, (MCBF). The equipment maintainability metrics are mean time to repair, (MTTR), measured in minutes and Preventive Maintenance, PM, measured in hours per year. Typical reliability and maintainability metrics for the interbay equipment are summarized in Table 33.29. The metrics are based on SEMI Standard E10 definitions and are in line with current ITRS [6] suggested values for best in class.

33.7 Anomaly Handling

Automated material handling system material handling anomalies are failures, functional errors, or deviations from normal operation. There are two general types of intervention: manual and automated. Manual intervention is required for most anomaly situations; however, some anomalies lend themselves to automated intervention and these can be addressed by some elements of AMHS. For example, OHV provide automated correction of some abnormal operations such as communications and carrier hoisting transfer retry attempts at an equipment load port. In general, automated intervention can serve to correct temporary or transient error situations and save the time of manual intervention in these cases.

Automated intervention implies the presence of software anomaly handling routines inside the control code that is automatically invoked when a certain set of pre-determined events occur and when the system operates in a degraded manner that is still viewed as safe. The system then tries to make repeated attempts to correct these forms of error and in most cases the new values over-ride the previous value.

TABLE 33.29 Reliability Metrics for Interbay and Intrabay Transport and Storage Equipment

Reliability Metric	Metric Definition	Requirement
Interbay transport system mean time to repair (MTTR)	Average time to correct the interbay transport system-related (vehicles, track, control system, etc) failure and return the system to a condition where it can perform its intended function without any degradation impact	10 min
Intrabay transport system MTTR	Average time to correct the intrabay transport system-related (vehicles, tracks, control system, etc) failure and return the system to a condition where it can perform its intended function without any degradation impact	10 min
Storage system MTTR	Average time to correct the storage (Stocker) system-related failure and return it to a condition where it can perform its intended function without any degradation impact	25 min
Interbay transport mean moves between failures (MCBF)	Average number of carrier move cycles (delivery from point A to point B) made by automated material handling system (AMHS) interbay transport equipment before a manual intervention is needed to fix a failure that causes the system to degrade performance or stop	9000 carrier moves
Intrabay transport MCBF (carrier)	Average number of carrier move cycles (delivery from point A to point B) made by AMHS intrabay transport equipment before a manual intervention is needed to fix a failure that causes the system to degrade performance or stop	9000 carrier moves
Storage system MCBF (carrier)	Average number of carrier move cycles (delivery from point A to point B) made by AMHS storage equipment before a manual intervention is needed to fix a failure that causes the system to degrade performance or stop	40,000 carrier moves

Manual intervention is mandatory in cases where functionality ceases and where personnel, payload, or equipment safety may be jeopardized. An example situation is when one of the sensors on the load port (presence sensor) indicates the presence of a FOUP on its load port while the other complementary load placement sensor does not detect the correct placement of the same FOUP on the same load port. When a condition occurs, where one sensor says everything is okay and another says something is not okay, (that is when the manual anomaly handling is really useful), a maintenance technician is needed to resolve the anomaly situation in a manual (non-automated) manner using a series of manual actions.

Automated material handling system user policies should be developed and put in place to guide operating staffs as to when their intervention is necessary; and how and by whom a manual intervention is to be made. Recovery of anomalies should be done by personnel with the appropriate level of training in order to provide for personnel and product safety; for proper recovery of the AMHS; logging of the intervention; and for diagnosis of the causes of the anomaly situation. As a minimum capability, AMHS elements generally come with equipment error logging and error code display features that serve to guide personnel making manual interventions as well as diagnosing the causes of anomaly situations.

33.8 Use of Computer Simulation for Designing and Operating AMHS

33.8.1 Benefits of Computer Simulation for AMHS Design and Operation

Use of computer simulation techniques is essential for the design and operational use of AMHS systems. Typically, the type of simulation used is based on discrete event simulation techniques. These simulators capture the dynamic behaviors of hundreds of production tools with thousands of work-in-process lots that are present in the modern IC manufacturing factories and integrate AMHS behaviors in these production simulations. Having these tools to design the configurations of AMHS systems is one of the most critical and essential tasks before undertaking the task of sizing the AMHS system. There are many reasons why this is beneficial. A few examples are listed as follows:

- When developing leading-edge fab processes and ramping them in high volumes, there are significant uncertainties and questions as to whether the factory systems that are designed upfront are capable of dealing with the resulting manufacturing variations that occur much later in the factory life-cycle. Dynamic, full-fab simulations have helped to answer many diverse business questions that are faced by manufacturing stakeholders without having to experiment with real wafers, tools or people in the actual factory. Sensitivity analysis using simulation models also enable factory customers to understand relative risks of potential options and to evaluate trade-offs to help them to make more data-based decisions on many uncertain areas of the business.
- Upfront understanding of AMHS performance and its impacts on fab production metrics is another benefit of simulation techniques. Simulation models are used extensively to understand the limitations of pervasive interbay and intrabay systems in 300-mm fabs. By integrating equipment capacity simulation models with AMHS models, a new ability is possible to understand different elements of fab cycle-time in a fully automated line. These kinds of models generate a useful Pareto of waiting times by category, providing a good tool for dialing in system performance and continuous improvement. It also enables manufacturing to prioritize improvements on those areas where the return or benefits are the highest.
- Estimating the number of stockers needed in the factory, the total storage requirements, and the number of interbay and intrabay vehicles needed to meet factory output can be ascertained with simulation tools well before the system is implemented. Full-factory simulation models are being used extensively to determine the number of stockers needed to meet the production output as well as the storage capacity of each stocker in the factory layout. The simulations permit the user to understand the variations of WIP levels in each area of the factory and that

data is used to size the stocker quantity and total storage requirements. Also, dynamic AMHS simulation of the transport systems enable the user to define the number of intrabay OHVs needed to ensure timely transportation of lots to each tool in the intrabay loop as well as the number of vehicles needed in the interbay loop to provide rapid transport between stockers in the AMHS layout. Simulation provides important clues on the throughput needs of the factory, which are typically expressed as average and peak moves per hour across each segment of the transport path as well as across the entire factory. The simulation models also generate average and peak delivery times for different vehicle count scenarios. Delivery time is defined as the time taken for lot transport from any one point in the factory to another defined point in the factory layout. The longer the delivery time, the higher is its impact on fab production. Fab design and operating staffs are always trying to minimize this metric. It can be seen that having all this information and insight into system performance well before the AMHS implementation is a very desirable capability to have.

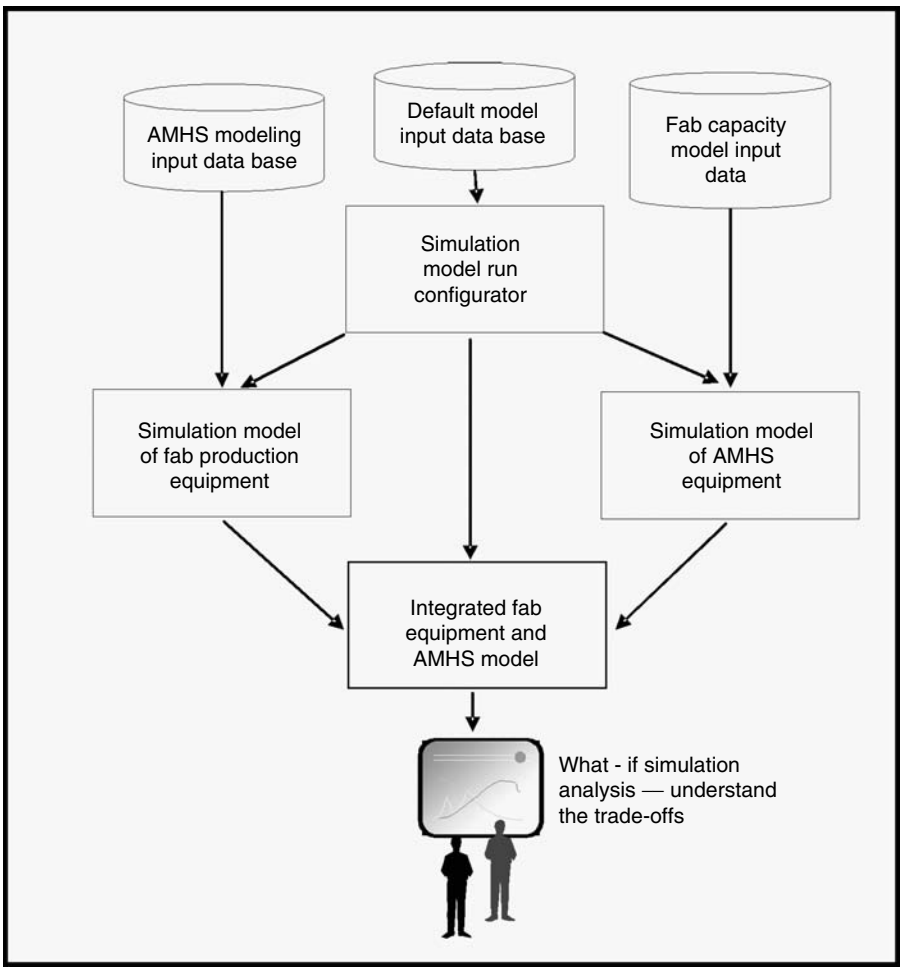

AMHS modeling input data base
Default model input data base
Fab capacity model input data
Simulation model run configurator
Simulation model of fab production equipment
Simulation model of AMHS equipment
Integrated fab equipment and AMHS model
What - if simulation analysis — understand the trade-offs

FIGURE 33.33 High level functional block for AMHS simulations integrated with fab production simulation tools.

33.8.2 Typical Computer Simulation Components

A very high-level functional block diagram is shown in Figure 33.33 for AMHS simulations integrated with fab production simulation tools. This configuration serves to generate an integrated simulation environment from which many important analyses can be performed to give the fab user some very important insights to AMHS performance.

33.8.3 How Simulation Is Used

Computer simulation of interbay and intrabay AMHS performance is employed to:

- Evaluate AMHS equipment configurations vs. the product handling traffic requirements such as throughput rates, delivery times, and "catch-up" capability requirements of the factory; here the goal is to configure the AMHS elements in order to support the manufacturing of the factory and to handle the production variabilities from one hour to the next.
- Estimate the required quantities of interbay and intrabay AMHS elements such as the number and operational requirements of transport vehicles, stockers (the number of stocker robots and input/output ports as well as the storage capacity of each stocker), and vertical lifters for those that lots have to be moved across different levels of the cleanroom;
- Do sensitivity analyses to define the constraints and limitations of transport track configurations and provide a means of testing alternative configurations until a satisfactory configuration is identified. In this exercise, all but one user defined parameter of the design are held fixed and the impact of the change on the overall system is evaluated over a series of small increments of the variable parameter.
- Support management of AMHS operations that support factory production over time; enable the users to comprehend how the system is currently used; forecast how future usage scenarios are likely to occur; and enable factory personnel to ensure the system performance of the AMHS meets production goals of the factory.
- Forecast requirements for upcoming additions and changes to the AMHS as the factory process flow(s) change overall time.

33.9 AMHS Implementation and Related Considerations

33.9.1 Implementation Phases and Responsibilities

The different phases of an example AMHS implementation are summarized in Table 33.30. These phases are applicable to implementation projects of any scope. The phase durations in Table 33.30 are for a 300 mm interbay and intrabay AMHS in all bays of a 20,000 wafer starts per month size fab. Some of the phases can overlap in time, which results in full fab AMHS implementations being accomplished in 12–18 months depending upon customer preparations, staffing, and schedule requirements, together with the capabilities the supplier(s). Multiple suppliers may be involved (sometimes simultaneously) in some fab implementations depending on the customer's assessments and AMHS configuration requirements.

33.9.2 Facilities Interfaces

When installing AMHS equipment in the cleanroom, it must be interfaced with several facility and building systems. Comprehending these interfaces and the interactions with these systems is essential for successful AMHS implementation. Table 33.31 summarizes the major interfaces and the reasons that they relate to each other.

Spatial location of intrabay transport systems including OHB must be comprehended early in the design of new facilities and AMHS applications. Table 33.32 presents some of the spatial and structural

TABLE 33.30 Example Automated Material Handling System (AMHS) Implementation Phases, Durations, and Responsibilities

Project Phases [Duration in Weeks]	Customer Responsibilities	Supplier Responsibilities
Requirements development [4–8]	Definition of the basic AMHS requirements including the expected fab manufacturing capacity; production and non-production material flows; facility architectural features; and an initial production tool layout. Preliminary sizing of work-in-progress (WIP) storage and transport requirements	Orientation of customer personnel on current state-of-the-art in AMHS equipment and control systems
Conceptual design [8–12]	Development of an AMHS conceptual layout design integrated with the architectural features of the fab facility and with the initial production tool layout	Supply of example AMHS equipment parameters and configurations as may be requested by customer personnel
Specifications and request for quotation (RFQ) [4–8]	Preparation of detailed specifications for each element of the AMHS as well as for the system as a whole including specifications for acceptance testing and warranty support. Transmittal of RFQ package to suppliers, which includes the fab requirements along with the AMHS conceptual layout, installation timeline, roles and responsibilities of the different parties during implementation and post install support, and specifications	Preparation of a response to the customer's RFQ package that includes: the expected performance metrics for the proposed AMHS configuration and each of its elements; a warranty support package; and a listing of items where exceptions are taken and clarifications are needed on the customer specifications. Estimation of supplier manpower needed to successfully install, qualify and turnover the system to the customer. Also estimation of post-install sustaining support, if needed
Purchase negotiations and supplier contracting [2–4]	Participation in AMHS technical and commercial negotiations with suppliers, evaluation of supplier RFQ responses, and selection of the supplier(s) for fab application. Award of the contract(s) to the chosen supplier(s)	Participation in AMHS technical and commercial negotiations as decided by the customer. Co-signing the contract if chosen as a supplier for the fab application
Detailed design and build [10–16]	Supply of any updated factory materials flow, layout, and facilities information to the supplier(s). Timely review, critique, and approval of supplier designs. Conduct of source inspections prior to shipment	Detailed design of equipment and control systems to meet the performance requirements of the customer fab followed by build of the equipment according to the agreed specifications and delivery schedule(s). Getting agreement on supplier manpower for installation, qualification, turnover and sustaining support (as needed)
Installation [10–18]	Supply of office and cut-shop spaces; local telephone support; equipment move-in; space for staging and erection of equipment coordination of supplier work efforts with other fab activities	Certification of personnel for work in customer fab. Erection, assembly, set-up, checkout, and exercising of the equipment in preparation for acceptance testing

(*continued*)

TABLE 33.30 ***(Continued)***

Project Phases [Duration in Weeks]	Customer Responsibilities	Supplier Responsibilities
Integration with tools and manufacturing execution system (MES) [6–8]	Scheduling and coordination of AMHS intrabay to production tools for loadport to overhead hoist vehicle (OHV) mechanical interface testing, E84 communications integration testing, and material control system (MCS) to MES integration efforts	Participation in AMHS intrabay to production tools mechanical interface testing including E84 communications integration as well as MCS to MES integration efforts
Testing of equipment and systems [2–4]	Supply of operations personnel, product carriers, instrumentation, and fab utilities together with coordination of AMHS testing with other fab operations	Operation of the equipment and systems in accordance with the acceptance test plans
Training [6–8]	Schedule time and travel as necessary for fab personnel to take formal training courses. Lead personnel to take part in equipment and systems set-up and check-out as part of on-the-job training. Lead personnel to train other fab personnel responsible for AMHS operation and support	Maintenance, operator, and systems operations training courses including training materials for AMHS and each of its various elements. On-the-job training for the lead maintenance and operations personnel of the fab
Production use	Operation and maintenance of AMHS in accordance with supplier recommended and required procedures. Implementation of AMHS related personnel and product safety policies as well as material logistics management procedures	Timely implementation of improvements and correction of all deficiencies identified during acceptance testing. Warranty and back up support timely notification of modifications for improvement of performance, functions, and reliability. Supply of all necessary spare parts in quantities to achieve AMHS MTBF, MTTR, and availability requirements
Additions and reconfigurations to existing installed system (as needed)	Establish a contractual agreement with supplier for support of additions and reconfigurations, on a case-by-case basis. In each case and on a smaller scope basis, provide definition of requirements and manage an implementation process as outlined herein	Provide on-going support for AMHS additions and reconfigurations on a project-by-project basis per prior contract agreements and per an implementation process similar to that outlined herein

parameters associated with OHT equipment. Due to the vertical dimensions of OHT equipment, ceiling heights in 300 mm automated fab tool areas tend to be in the range of 4350–4700 mm above the cleanroom flooring.

33.9.3 Guidance for Installation, Testing, and Training

Table 33.33 and Table 33.34 summarize installation tasks, resources, and tests, which lead to successful installation and acceptance of AMHS. When possible, it is preferable to move in and install AMHS equipment before the production equipment due to the large amounts of space required to stage and erect AMHS equipment. Planning and timely "follow-through" on task and resource commitments are essential for meeting time schedules. Schedule risks are reduced in installation efforts if experienced staff

TABLE 33.31 Summary of Facilities Interfaces with Elements of Automated Material Handling System (AMHS)

Facility Features and AMHS Elements	Floors and Sub-Floors	Ceilings (12 ft+ Clearance Elevation)	Walls and Partitions	Air Handling and Environment	Safety, Personnel and Fire	Utilities, Primary
Stockers	Structural support and bases	Vertical clearance	Clearance for maintenance access	Ducted air or power for fan filter units	Fire sprinklers	Electrical and pneumatics
Overhead tracks	Vertical clearance; locations for power supplies	Hanging loads and fixtures at ceiling; vertical clearance above tools	Clearance for vehicles	Vertical laminar flow (VLF) air; electro-static discharge ionizers	Avoidance of conflicts with fire sprinklers; automated fire door for fire wall pass-through	Electrical power supplies, their locations, and routing of cables to transport tracks
Automated guided vehicle (AGV's), rail guided vehicles (RGV's)	Floor surface and support	VLF air	Safety and pinch points	VLF air	Egress aisles/widths and safety panels	Electrical power
Inter-level lift	Passage easements and maintenance space	Clean air	Air flow balance and channeling in lift shaft	Air flow and pressure balancing	Fire door for fire barrier; access interlocks	Electrical power
Carrier tracking systems		Install of "line of sight" transceivers and associated wiring	Obstructions to line of sight systems			Isolated electrical power
Communication networks	Cable conduits and routing	Cables and radio frequency (RF) transceiver access points		RF communication		Isolated electrical power
Control systems	Locations for work stations			Computer cooling	Monitor, detection, alarm signal	Isolated electrical power

TABLE 33.32 Example Factory Interface Metrics for Overhead Hoist Transport (OHT) Equipment

OHT Single Level System Elements and Facilities Features	Heights of OHT Elements (mm)	Alignment to Undocked Facial Datum of Tool Load Port or Nest (±mm)	Track to Track Centerlines Spacing (mm)	Clearance Between Track Centerline and Wall (mm)	Maximum Hanging Loads of 300 mm OHT+FOUPs at Ceiling Attach Points (kg)	
					1200 mm Hanger Pitch	1800 mm Hanger Pitch
Track and vehicles	980–1200	3	≥500	≥350	120	180
Overhead buffer (OHB)	460–520 and required floor clearance of > 2600	3	≥500	≥350	20	25

TABLE 33.33 Installation Considerations

Installation Related Items	Considerations and Typical Practice
Customer related prerequisites	Equipment locations and the order of equipment install should be predetermined with the supplier along with the following: cleanroom areas completed with ceiling grid and flooring installed; ceiling filters installed, scanned, and certification complete; equipment receiving dock available and move-in route(s) provided; floor locations marked, support bases ready, and utility ports cut as needed; and staging areas marked and lay down spaces defined
Supplier related prerequisites	Equipment should be clean, complete, and delivered per preplanned schedule. Sufficient, skilled installation staff should be in place; fab cleanroom certified; and trained to provide for efficient equipment installation, integration, and test support
Sequence of equipment installation (typical)	Interbay track Stockers Intrabay track or unified track Intermediate level system controllers Material control system (MCS)
Timing of installations	It is desirable to install as much of the interbay and intrabay equipment as possible prior to production tool installations. This eliminates equipment move-in congestion issues as well as providing for more timely automated material handling system (AMHS) integration and production ramp support
Equipment unit level "burn-in" and cycle testing	Equipment unit level "burn-in" or functional cycling serves to identify where adjustments are needed and to demonstrate operational functionality and stability prior to unit level testing and system integration
Unit level testing	Testing of individual units of the AMHS serves to insure a higher level of success when individual units are integrated together as a system
System integration (and testing)	System integration involves: (1) functional bringing together of the equipment units and their control systems with the MCS and host factory control system; (2) operating and testing the equipment as a system; (3) making refinements to the AMHS operation; and in the case of intrabay AMHS, integrating the AMHS operations and communications with the production tools
System testing	System level testing results in the testing of the system of equipment when it is under the control of the MCS and factory host control system. System level testing can proceed as a series of sub-system tests followed by tests of the overall system. The tests should measure performance metrics, which include: reliability; throughput capacity; carrier deliver times; and carrier handling quality, and error recoveries
Expansions and reconfigurations	AMHS operational downtime of ≤60 min should be achieved. The keys to minimizing system downtime during an expansion or reconfiguration are planning and preparation in advance. All installation work and testing of the new elements should be done in advance of connecting them with the existing production equipment. All materials and resources should be in place for the actual connection to the existing equipment and systems

TABLE 33.34 Example Listing of Automated Material Handling System (AMHS) Acceptance Criteria and Tests

No	Acceptance Criteria and Tests	Eqpt. Type	Equipment Quantity	Test Locations	Number of Measures	Corrective Actions
1	General construction and quality					
2	Safety compliances per SEMI S2 and user requirements					
3	Unit cycle tests					
4	Equipment functions					
5	Equipment vibration and audible noise					
6	Quality of product handling					
7	Cleanliness					
8	Electrostatic charge					
9	Communication systems					
10	Control systems, user interfaces, and software					
11	Inter-equipment transport					
12	Tool-to-tool transport					
13	Anomaly handling					
14	Systems throughput and delivery times					
15	Correction of any shortcomings					

are involved in key positions; detailed inventories are made of equipment components and supplies upon receipt to determine what may be missing (so they can be re-ordered ahead of their need dates); the proper mix of hardware and software staff members are present during the integration phases of the installation; and appropriate quantities of key spare parts are provided to support the AMHS installation and acceptance testing. Both customer and supplier staffs should complete as much of their work as possible prior to delivery of the AMHS equipment on site. There is no substitute for teamwork and recognition of the inter-dependence of customer and supplier staff roles in achieving efficient, successful AMHS installations.

TABLE 33.35 Example Worksheet for Automated Material Handling System (AMHS) Training Tasks and Resources

AMHS Training Tasks/Resources	Person(s) Responsible	Resources Needed	Planned Completion Date	Actual Completion Date
Training of equipment technicians on maintenance procedures, diagnostic methods, on site repairs, and anomaly handling				
Training of material control system (MCS) staff				
Establish communication support with MCS supplier site				
Training of lead manufacturing staff as trainers for fab operators				
Training of fab operators on use of and interaction with the AMHS				
Set up of preventative maintenance (PM) schedules				
Inventory and management of equipment spare parts				
Scheduling and conduct of preventative maintenance per supplier requirements				

TABLE 33.36 Extendibility and Scalability Implications of Automated Material Handling System (AMHS)

Sub-System Impacted by the Change	Typical Changes to Configurations that Must be Expected Over the Lifetime Once Installed	Extendibility, Scalability and Flexibility Features that Must be Comprehended Upfront During Initial System
Factory layout changes	The size of the factory may expand or contract depending upon business needs. In some cases, a few additional bays may get added. In other cases, some adjacent bays may get combined to form a larger bay. Sometimes, an adjacent factory (area) could get merged with the current layout to form a much larger factory, and operating as one large mega factory	In all these cases, it is imperative that the AMHS design be as modular as possible so that future expansion to the original AMHS design is possible vs. having to completely de-install and re-install each time the factory changes. Particular attention must be paid to how the transport loops are configured towards the edges of the bays and the main aisles. Provisions must be made so that future track joining and merging is possible. Most importantly, it is very important to keep the cleanroom levels of the new expansions at the same level as the original cleanroom, as it could significantly impact both work-in-progress (WIP) transport capability as well as people movement between these areas
Factory throughput changes	In integrated circuit (IC) manufacturing, each new process technology results in additional processing steps. Consequently, for the same wafer start rate, and also driven by productivity improvements over time, the throughout requirements of the factory could increase over time	The moves per hour of the AMHS system (both the interbay and the intrabay systems) are most likely going to increase over time. The impact is most felt if the increase is limited to a section of the factory layout vs.equal increase across all functional areas. This implies that the designed moves rate headroom that must be provided upfront is in those areas where the current wafer starts rate is already most constrained
Factory storage changes	Storage requirements are likely to increase over time as new processes are added and the number of non-production wafers is increased for engineering troubleshooting reasons during the early stages of the new process insertion	Leave white spaces in the layout for future stocker placement. Design the stocker storage requirements upfront to include these unanticipated needs in the future. Ensure that the track routing for interbay and intrabay systems are capable of accessing these new stockers without causing transport bottlenecks or throughput constraints
WIP Stocker configuration changes	Some stocker designs permit additional storage shelves to be added by growing the stocker in the linear direction or growing the stocker taller	The stocker expansion must comprehend interbay and intrabay input/output port compatibility with the currently installed track heights as well as the precise locations where transport vehicles interface with the stocker for carrier load and unload. Sometimes, the stocker robot is speeded up for increasing the throughput of the stocker. The time to de-install the older stocker robot and re-install the enhanced robot must be minimized so that the downtime impact to the factory is negligible
Interbay and intrabay tracks configuration changes	As the layout expands by adding more bays to the layout or annexing layout areas adjacent to the current factory, the interbay tracks will see significant changes such as track extension, track merging and comprehending elevation differences	The interbay transport track design and components must be modular so that rail joining and extending can be quickly achieved. The transport supplier should also provide a lightweight, high throughput lifter solution that enables WIP movements between rails that are at different levels. This will provide layout designer the flexibility of making only minor changes to the older systems that need to be merged or connected

(*continued*)

TABLE 33.36 ***(Continued)***

Sub-System Impacted by the Change	Typical Changes to Configurations that Must be Expected Over the Lifetime Once Installed	Extendibility, Scalability and Flexibility Features that Must be Comprehended Upfront During Initial System
Reticle storage system changes	Sometimes it may not be possible for the IC maker to store all reticles in only one standardized carrier such as the reticle standard mechanical interface (SMIF) pod (RSP). Other non-standard carriers may need to be integrated into the system	Two or three different reticle storage systems may be needed if the reticle carriers to be stored are significantly different dimensionally from each other. Sometimes reticle transport may be possible only for one reticle carrier type, while a decision may be made not to provide for storage for the non-standard carrier type
Material control system (MCS) changes	Addition of new process technologies, changing names and capacities of Stockers, and bringing the new AMHS system back to production in the shortest possible time in order to minimize the downtime of the overall system	The MCS must be capable of adding new AMHS elements and production tools into the database very rapidly and without impacting on-going production operations. For example, it should be possible to add a new stocker to the existing AMHS and to quickly define the track IDs that are connected to it as well as make the storage size a variable so that the addition of new shelves can be quickly comprehended. Once the above tasks are accomplished, the MCS must be capable of automatically computing the shortest from/to path based on the new transport configurations. If the from/to requires traversing across multiple transport systems and stockers, the MCS must be capable of auto-calculating the shortest or most efficient transport route that must be taken

In addition to successful completion of final acceptance testing, start up of AMHS requires training of fab support and operations staffs; certification of AMHS equipment by fab quality assurance; and acceptance for production use by manufacturing management. Training of fab staffs should be planned and performed by AMHS equipment and MCS supplier personnel prior to, during, and following AMHS installation. Fab operator training may be accomplished by having supplier staff, train fab "lead" manufacturing and AMHS persons, who in turn train the operators on their respective work shifts. An example worksheet for training tasks is presented in Table 33.35.

33.10 Extendibility and Scalability of AMH Systems

Extendibility and scalability are important considerations for AMH systems since the production requirements of established wafer fabs change over time. The AMH equipment and control systems need to be designed with these considerations in mind. In addition, anticipation of wafer fab changes is needed in the planning of all AMHS applications. Table 33.36 presents a summary of example wafer fab changes, resulting changes to AMHS configurations, and features of AMHS that should be applied to achieve extendibility and scalability.

Appendix

SEMI AUX 006 User Interface Configurations for 300 mm Equipment Load Ports
SEMI E1.9 Provisional Mechanical Specification for Cassettes Used to Transport and Store 300 mm Wafers

SEMI E4	Equipment Communications Standard 1, Message Transfer (SECS-I)
SEMI E5	Equipment Communications Standard 2, Message Content (SECS-II)
SEMI E10	Standard for Definition and Measurement of Equipment Reliability, Availability, and Maintainability (RAM)
SEMI E14	Measurement of Particle Contamination Contributed to the Product from the Process or Support Tool
SEMI E15	Specification for Tool Load Port
SEMI E15.1	Provisional Specification for 300 mm Tool Load Port
SEMI E19.4	SMIF
SEMI E23	Specification for Cassette Transfer Parallel I/O Interface
SEMI E30	Generic Model for Communications and Control of SEMI Equipment (GEM)
SEMI E33	Specification for Semiconductor Manufacturing Facility Electromagnetic Compatibility
SEMI E35	Cost of Ownership for Semiconductor Manufacturing Equipment Metrics
SEMI E37	High-Speed SECS Message Services (HSMS) Generic Services
SEMI E43	Recommended Practice for Measuring Static Charge on Objects and Surfaces
SEMI E47	Specification for 150-mm/200 mm Pod Handles
SEMI E47.1	Provisional Mechanical Specification for Boxes and Pods Used to Transport and Store 300 mm Wafers
SEMI E54	Sensor/Actuator Network Standard
SEMI E58	Automated Reliability, Availability, and Maintainability Standard (ARAMS): Concepts, Behavior, and Services
SEMI E57	Provisional Mechanical Specification for Kinematic Couplings Used to Align and Support 300 mm Wafer Carriers
SEMI E62	Provisional Specification for 300 mm FIMS
SEMI E63	Provisional Specification for 300 mm Box Opener/Loader to Tool Standard (BOLTS) Interface
SEMI E64	Provisional Specification for 300 mm Cart to SEMI E15.1 Docking Interface Port
SEMI E72	Provisional Specification and Guide for 300 mm Equipment, Footprint, Height, and Weight
SEMI E75	Provisional Mechanical Specification for Box/Pod Compatible Cassettes Used to Transport and Store 300 mm Wafers
SEMI E78	Electrostatic Compatibility—Guide to Access and Control Electrostatic Discharge (ESD) and Electrostatic Attraction (ESA) for Equipment
SEMI E82	Specification for Interbay/Intrabay AMHS SEM (IBSEM)
SEMI E84	Specification for Enhanced Carrier Handoff Parallel I/O Interface
SEMI E85	Specification for Physical AMHS Stocker to Interbay Transport System Interoperability
SEMI E87	Specification for Carrier Management (CMS)
SEMI E109	Provisional Specification for Reticle and Pod Management (RPMS)
SEMI E109.1	Provisional Specification for SECS II Protocol for Reticle and Pod Management (RPMS)
SEMI E111	Provisional Mechanical Specification for a 150 mm Reticle SMIF Pod (RSP150) used to Transport and Store a 6 In. Reticle
SEMI E112	Provisional Mechanical Specification for a 150 mm Multiple Reticle SMIF Pod (MRSP150) used to Transport and Store a 6 In. Reticles
SEMI E117	Provisional Specification for Reticle Load Port
SEMI E119	Provisional Mechanical Specification for Reduced Pitch Front Opening Box for Interfactory Transport of 300 mm Wafers
SEMI E129	Guide to Assess and Control the Electrostatic Charge in a Semiconductor Manufacturing Facility
SEMI F47	Specification for Semiconductor Processing Equipment Voltage Sag Immunity

SEMI M1	Specifications for Polished Monocrystalline Silicon Wafers
SEMI M1.15	Standard for 300 mm Polished Monocrystalline Silicon Wafers (Notched)
SEMI M8	Specification for Polished Monocrystalline Silicon Test Wafers
SEMI M8.12	Standard for 300 mm Polished Monocrystalline Silicon Test and Monitor Wafers (Notched)
SEMI M11	Specifications for Silicon Epitaxial Wafers for IC Applications
SEMI M12	Specification for Serial Alpha-numeric Marking of Front Surface of Wafers
SEMI M13	Specification for Alpha-numeric Marking of Silicon Wafers
SEMI M28	Specification for Developmental 300 mm Diameter Polished Single Crystal Silicon Wafers
EMI M29	Specification for 300 mm shipping box
SEMI M31	Provisional Mechanical Specification for Front-Opening Shipping Box Used to Transport and Ship 300 mm Wafers
SEMI P37	Specification for Extreme Ultraviolet Lithography Mask Substrate
SEMI S2	Environmental, Health, and Safety Guidelines for Semiconductor Manufacturing Equipment
SEMI S8	Safety Guideline for Ergonomics Engineering of Semiconductor Manufacturing Equipment
SEMI T1	Specification for Back Surface Bar Code Marking of Silicon Wafers
SEMI T7	Specification for Back Surface Marking of Double-Side Polished Wafers with a 2-D Matrix Code Symbol
SEMI T11	Specification for Marking Hard Surface Reticle Substrates (Data Matrix per ISO/IEC 16022)

References

1. Doering, R., and Y. Nishi. *Handbook of Semiconductor Manufacturing Technology*, 1st ed., New York: Marcel Dekker, Inc., 2000.
2. SEMI Standards, Semiconductor Equipment and Materials Inc., San Jose, 2005.
3. Ferrell, J., and M. Pratt. I300I Factory Guidelines: Version 5.0, Technology Transfer Doc. #97063311G-ENG, International SEMATECH, Austin, 2000.
4. Goodall, R. I300I Guidelines on 300 mm Process Tool Mechanical Interfaces for Wafer Lot Delivery, Buffering, and Loading, Rev. D, Technology Transfer Doc. #97063298A-XFR, International SEMTECH, Austin, 1996.
5. CIM Global Joint Guidance for 300 mm Semiconductor Factories, Release 5, Japan 300 mm Semiconductor Technology Conference, Japan 300 and International SEMATECH, Austin, 2000.
6. The International Technology Roadmap for Semiconductors (ITRS), 2004 edition, SIA Semiconductor Industry Association, 2006, http://www.itrs.net
7. Christal, L., and S. Fulton. 300 mm Best-Known Practices (300 BKP) for 300 mm Factory Integration, Technology Transfer Doc. #00124063B-ENG, International SEMATECH, Austin, 2001.

Further Reading

2nd Lecture, IC Factory Design for 300 mm Wafer Line Standardizing Study, Japan 300 mm Semiconductor Technology Conference (J300), Waseda University, Tokyo, 1996.

300 mm Physical Interfaces and Carriers SEAJ/SEMI Joint Seminar, 2nd ed., February 1997, Tokyo, 1997.

Bass, E., and P. Jai, Metrics for 300 mm Automated Material Handling Systems (AMHS) and Production Equipment Interfaces: Revision 1.0, Technology Transfer Doc. #97123416B-TR, International SEMATECH, Austin, 1998.

Dhudshia, V. H. *Hi-Tech Equipment Reliability: A Practical Guide for Engineers and the Engineering Manager*. Sunnyvale: Lanchester Press Inc., 1995.

Foster, L., and R. Doering. "Future Fabs: Manufacturing with 300 mm Wafers." *Technol. J.* 13 (1996): 5 (Texas Instruments, Dallas).

Garbayo, J., and R. Missale. "Automation for 300 mm Semiconductor Fab Operations." *Future Fab Int.* 6, (1998).

Gargini, P., and D. Pillai. "Factory Considerations for High Volume Manufacturing Using 300 mm Wafers." In *Proceedings of SEMI Hosted SEMICON Japan Symposium on 300 mm Wafer Fab Designs*, Makhuhari, 1997.

Gartland, K., and S. Kono. Automated Material Handling System (AMHS) Framework Document: Version 1.0, Technology Transfer Doc. #99073793A-TR, International SEMATECH, Austin, 1999.

Ghezzi, G., and L. Christal. Guidelines for the Installation and Alignment of 300 mm Overhead Transport Systems (OTS) and Load Port Interfaces, Technology Transfer Doc. #02064276A-ENG, International SEMATECH, Austin, 2001.

Harris, C. M., and C. E. Crede. *Shock and Vibration Handbook*. 2nd ed. New York: McGraw-Hill Book Co., 1976.

Janakiram, M., and J. S. Pettinato. "An ITRS Perspective on 300 mm Factory Advances and a Vision for Future Breakthrough Enterprise Performance." *Future Fab Int.* 18, (2005).

Johnson, R. W. "Design Rules for Fab CIM." *Solid State Technol.* September, (1996).

Liao, D., and H. Fu. "Speedy Delivery: A Simulation-Based, Two-Phase Approach for Dynamic OHT Allocation and Dispatching in Large-Scale, 300 mm AMHS Management." *IEEE Robotics and Automation Magazine* 11 (2004): 4.

Missale, R. "Expectations on Automation for 300 mm Semiconductor Fab Operations." *Future Fab Int.* 5, (1998).

Mitchell, P. E. "Automated Material Handling Systems.", *Tool and Manufacturing Engineers Handbook*, Vol. 9, Michigan: Society of Manufacturing Engineers, 1998, chap. 13.

Mulcahy, D. E. "Economic Justification Process." In *Materials Handling Handbook*, New York: McGraw-Hill Co., 1999.

Murray, A. M., and D. J. Miller. "Automated Reticle Handling: A Comparison of Distributed and Centralized Reticle Storage and Transport." In *Proceedings of 2003 Winter Simulation Conference*, New Orleans, 2003.

Pillai, D. "Material Handling Automation for Wafer Fabrication Facilities." In *Proceedings of IEEE/CHMT Conference and Symposium*, 277. Cambridge, MA, 1990.

Pillai, D. "Designing Automated Material Handling Systems for Large Scale Wafer Fabrication Automation." In *Proceedings of Society of Manufacturing Engineer's Conference on Semiconductor Manufacturing*, Phoenix: Society of Manufacturing Engineer's, 1989.

Pillai, D., and P. Calame. "Systems Integration and Automation for Sub-Micron Wafer Fabrication." In *Proceedings of Society of Manufacturing Engineer's Conference on Automated Clean Room Processes*, Orlando: Society of Manufacturing Engineer's, 1989.

Pillai, D., et al. "Integration of 300 mm Fab Layouts and Automated Material Handling Systems." In *International Symposium on Semiconductor Manufacturing (ISSM) Conference and Proceedings*, San Jose, 1999.

Pillai, D., and S. Srinivasan. "Material Handling Automation—Trends, Vision, and Future Plans." In *Conference and Proceedings of International Symposium on Semiconductor Manufacturing (ISSM)*, Tokyo, Japan, 1996.

Pillai, D., et al. "300 mm Full Factory Dynamic Simulations for 90 nm and 65 nm IC Manufacturing." *IEEE Transactions on Semiconductor Manufacturing*, special invited paper, IEEE, 2004.

Robertson, F., and L. Foster. "300 mm Factory System Integration." *Future Fab Int.*4 (1998).

Robertson, F., and P. Gargini. "300 mm Conversion Capability." *Future Fab Int.* 5, (1998).

34

Factory Modeling

34.1 Objectives: How Is Factory Modeling Used? **34**-1
Performance Metrics and Key Relationships • Fab Design • Operations Policies

34.2 Static Modeling **34**-10
Capacity and Long-Run Utilization Profiles • Cell and Cluster Tool Models • Cost of Ownership • Static Financial Projections

34.3 Dynamic Modeling **34**-16
Monte Carlo Discrete-Event Simulation • Queuing Models of Networks

Acknowledgments **34**-21

References **34**-22

Samuel C. Wood
Responsive Learning Technologies

34.1 Objectives: How Is Factory Modeling Used?

Factory modeling is used to predict the effect of decisions on factory performance, before incurring the risk or cost of actually implementing the decisions. The rate at which wafers can be processed in a fab, the average work-in-process (WIP) wager inventory, and the average cost per wafer are examples of factory performance metrics. Section 34.1.1 of this chapter describes the most common performance metrics of interest to factory modelers. The decisions that models are used to evaluate are divided into two groups: Section 34.1.2 describes typical fab design decisions such as tool choice, tool count, and cluster tool configuration; and Section 34.1.3 describes typical operational policies such as how lots are released into a fab, and the order in which lots are processed at a given tool.

34.1.1 Performance Metrics and Key Relationships

34.1.1.1 Throughput, Capacity, and Utilization

The most commonly modeled performance metric is throughput (also called "throughput rate"). Throughput is the amount of work processed by a set of resources over some time period:

$$\text{Throughput} = \frac{\text{Work completed during a time period}}{\text{Length of the time period}} \tag{34.1}$$

Work may be expressed as lots, wafers, or yielded chips, for example. Typical units of time are hours, days, or months. For example, the throughput of a process tool during a given week could be expressed in wafers per hour, and could be calculated as the amount of wafers that the tool processed, divided by the number of hours that the tool was available for processing. The long-run average throughput of a factory could be expressed in wafers per month, and could be calculated as the number of wafers completed over

a long period of time, divided by the number of months in that period of time. Factory models should define "work" and "time" as precisely as possible. For example, if throughput is expressed as "wafers per month," do "wafers" refer to product wafers and test wafers, or just product wafers that were not scrapped? Does "month" refer to an entire month, or just fraction of the month that the factory was actually operating?

There are two common, different definitions of capacity. *Throughput capacity* is the maximum sustainable throughput of a resource or a group of resources. Like throughput, this type of capacity is measured in work per time period (e.g., wafers per hour). Typical resources that would have associated throughput capacities are operators, processing tools, or entire factories. *Wafer capacity* is the maximum number of wafers that can be stored at an inventory point or simultaneously processed in a tool. For example, if a furnace runs a single process on a maximum of 100 wafers at a time, and the entire time to run the process (including loading and temperature ramping) is 5 h, then the throughput capacity of the furnace is 100/5 = 20 wafers per hours, and the wafer capacity of the furnace is 100 wafers. The term "capacity" is often used in isolation, and the type of capacity must be determined from context. As with throughput, "work" and "time" must be precisely defined when specifying throughput capacity. In the above furnace example, the unit of time was actually hours that the furnace was available for processing. Instead if the unit of time were clock hours, including time that the furnace is down for maintenance or repair, then the throughput capacity would have been lower.

The throughput capacity of a system of multiple resources, such as a bay or a fab, may often be limited by one or more of the resources in the system. Resources that constrain the throughput capacity of a system are called *bottlenecks.* For example, if a fab running a single process consists of 40 types of tools, with several copies of each type, and there are always operators and material handling equipment available to feed the tools, then the fab bottleneck would be the type of tool with the smallest throughput capacity. In general however, the throughput capacity of a fab, and the fab bottlenecks, are not so clearly defined. For example, if the fab is running multiple process recipes, then the fan capacity and bottlenecks will probably depend on the mix of the different process recipes being run in the fab. If one recipe requires more aluminum etching steps than another recipe, then the aluminum etcher may only become a bottleneck if most of the wafers in the fab are processed using the recipe with the larger number of aluminum etching steps.

The throughput capacity of a system depends on its various physical characteristics, including: the number of operators, tools, and other resources; the reliability, flexibility, and set up requirements of process tools; processes and recipes that determine bottleneck resources, yields, and rework requirements; and ability to quickly perform set ups, processes, maintenance, and repairs. Throughput also depends on operating policies, which will be described in Section 34.1.3.

Utilization is the fraction of time that a resource such as a process tool or operator is processing work. The actual meaning of utilization is vague, unless the activities that constitute "processing" and the time units are carefully defined. For example, processing may refer to only the time a tool is processing product wafers, or it may also include loading and unloading times and the time processing test wafers. The unit of time may just be the hours that the tool is available for processing, or it may be all 24 h per day. A utilization profile is a list of the fractions of time that a resource spends in different states and performing different activities, such as processing wafers, undergoing maintenance, or waiting for new material. Semiconductor equipment and materials international (SEMI) standard E-10 [17] defines a detailed hierarchy of such activities for processing tools. As an alternative to utilization profiles, overall equipment efficiency (OEE) describes tool productivity as a product of factors such as the fraction of time a tool is available and the fraction of available time a tool is processing. OEE is described in SEMI standard E-79 [20].

34.1.1.2 Cycle Time

Another common factory performance metric is *cycle time,* also called throughput time. Cycle time is the length of time that passes while some specified set of tasks are performed. For example, fab cycle time typically refers to the average time a wafer spends in the fab, from when the raw wafer is released into the

fab until the completed wafer exits the fab. Average fab cycle times are determined by averaging the cycle time of each lot, over all the lots completed over some time period. Similarly, the cycle times of a single processing step or sequence of steps is the time that elapses from when a lot begins the step or sequence, until the lot completes the step or sequence.

Feb cycle time includes both queuing time and processing time components. The queuing time is the total length of time that a lot spends waiting to be processed on a tool. The lot may have to wait because the tool is processing other lots, is undergoing a setup or repair, or is waiting for an operator to load the lot. Although the definition of processing time seems obvious, it may be unclear whether the process time in a model is defined to include times from sources such as tool loading and unloading or lot transportation. Fab cycle time is frequently specified as a multiple of processing time, raw cycle time, or ideal cycle time. As with the definition of process time, definitions of raw or ideal cycle time also vary considerably from company to company. Typical high-volume fab cycle times are reported to be 4–10 times process time, which means that lots spend most of their time waiting for tools instead of being processed. Reasons for these long queue times are discussed in Section 34.1.1.5. The ratio of cycle time to number of masks is a useful metric for comparing cycle times for different process recipes or different fabs. Typical contemporary fabs experience cycle time of 1.5–5 days per masking level [10].

34.1.1.3 Inventory

Raw materials inventory refers to the amount of raw materials that have not yet undergone processing, *WIP inventory* is the amount of materials that are undergoing processing, and *finished-goods inventory* is the amount of materials that have completed processing. Fab modeling generally focuses the WIP inventory of wafers in the fab. The average WIP inventory level is the number of wafers that have been released into the fab but have not yet completed processing, averaged over time.

The cost of inventory is typically divided into financial holding costs and physical holding costs. A certain amount of money is required to sustain WIP inventory. For example, if raw wafers cost \$150 a piece and there is a WIP of 10,000 wafers, then the value of the WIP includes a raw wafer materials value of $\$150 \times 10{,}000 = \$1{,}500{,}000$. The financial holding cost of inventory is the return that could have been earned on that money if it was invested somewhere else. Physical holding costs include the cost to insure the WIP inventory, and the cost of wafers becoming obsolete before they could be completed. Many companies group physical and financial holding costs together as some total rate. For example, if the inventory holding cost is 25% per year, then the holding cost of each \$1 of inventory value is \$0.25 per year. Holding costs of 20%–35% per year are commonly used.

34.1.1.4 Little's Law

Little's Law [11] describes the relationship between throughput, cycle time, and inventory:

$$L = \lambda W \tag{34.2}$$

where L is the average WIP inventory level, λ is the average throughput, and W is the average cycle time. Little's Law applies to any system, as long as the total number of jobs that entered the system is equal to the total number of jobs that exited the system over the time period being considered. For example, if the system is a fab that experienced 1000 lot starts per month over a year, and over that same year the average cycle time was 2 months, then Little's Law predicts that the average WIP inventory would be $1000 \times 2 = 2000$ lots. This estimate is valid to the extent that the total number of lots started during that year was approximately equal to the total number of lots that was completed during the year.

34.1.1.5 Congestion and Variability

To minimize cycle time and inventory, it would be ideal for each lot to arrive at each process tool just as that tool became ready to process that lot. In such a scenario, tool would experience high utilization,

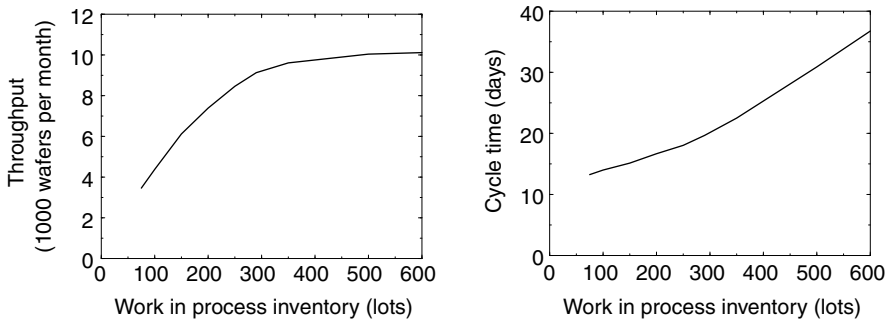


FIGURE 34.1 Results from a simulation of a fab with a capacity of 10,000 wafers per month. Each plotted point was generated by a separate simulation where the work-in-process was held constant by introducing a new lot when another lot was completed.

while cycle times and WIP inventory levels would be kept low. The experience in actual fabs is quite different however. As mentioned above, a typical lot spends the large majority of its time in most fabs waiting rather than being processed.

Figure 34.1 shows a typical relationship between long-run average throughput, average cycle time, and average WIP inventory. (The data came from simulations of a fab with a throughput capacity of 10,000 wafers per month and a lot size of 20 wafers.) The plot on the left shows that a large amount of WIP inventory is required to achieve a high throughput. For throughput values that are close to capacity, a relatively large increment in WIP inventory corresponds to a relatively small increase in throughput. The plot on the right shows how cycle time increases as WIP inventory increases.

As the number of lots in the fab increases, a given lot is more likely to arrive at a machine that is busy processing some other lot. Thus, the waiting time of each lot increases with increasing WIP inventory.

A large WIP is required to sustain a high throughput because of *congestion.* Congestion disrupts the smooth flow of lots through the fab, preventing lots from arriving only when tools are ready for them. Congestion can be caused by variability from tool maintenance and random failures and from random delays that result when an operator is temporarily unavailable to load a lot into a tool. The random introduction of now lots into the fab is another source of variability that contributes to congestion. Congestion also results when batch tools such as furnaces dump several lots in front of single-lot processing tools that must then work through the burst of work. Finally, the re-entrant characteristic of semiconductor process flows contributes to congestion. A re-entrant process flow is a flow where the same tool group (e.g., the deep-UV steeper group) is visited more than once over the course of the process. A re-entrant flow makes it especially difficult to minimize lot waiting times by coordinating processing and material flows. Efforts to reduce congestion can be grouped into fan design (described in Section 34.1.2) and operations policies (descried in Section 34.1.3).

34.1.1.6 Transient vs. Steady–Steady Performance

Figure 34.2 shows the results to simulating a fab for 40 weeks. (Simulations are further described in Section 34.3.1). At the beginning of the simulation, 400 lots were released into the empty fab, and subsequently, a new lot was released into the fab each time a lot was completed, to maintain the total WIP inventory at 400 lots. The number of wafers completed each week is plotted in the left graph, and the cycle time of each lot, in the order completed, is plotted in the right graph. Roughly the first 10 weeks or the first 800 lots show the *transient response* of the fab: as wafers distribute themselves in the fab, the throughput climbs and the cycle time first rise and then falls. The subsequent weeks and lots show a *steady-state response:* even though throughput and cycle time continue to vary, the average and variability of their values seem relatively stationary. Both the transient and steady-state performance of

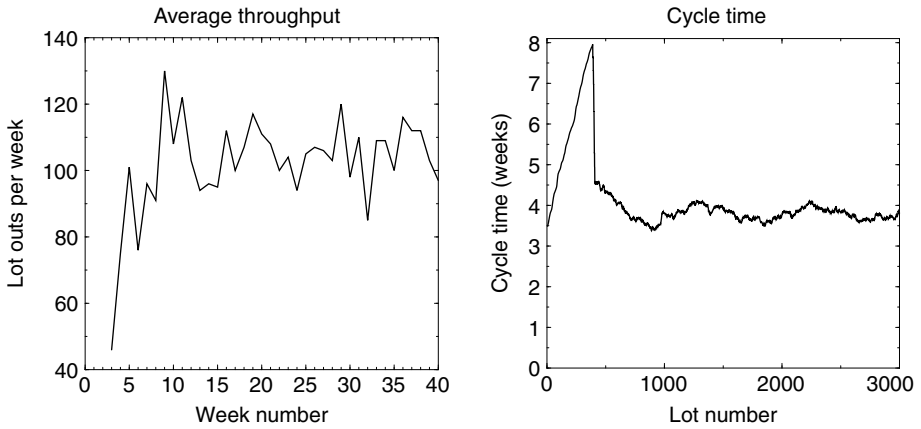


FIGURE 34.2 The evolution of average throughput and cycle time during a single fab simulation.

a factory depend on the factory design and operating policies. However, when modeling fab performance, it is important to specify whether the performance metrics are transient or steady state. Describing steady-state performance is easier and more common, because in the steady state, it is meaningful to describe performance metrics by their long-run average values or distributions.

34.1.2 Fab Design

A common objective of factory modeling is the design of a fab that meets specified performance requirements. The performance requirements almost always include a throughput capacity. There may also be cost objectives such as "minimize the total cost of ownership" (CoO) or cycle time objectives such as "the 95th percentile cycle time should not exceed 6 weeks." Fab modeling using the techniques descried in Section 34.2 and Section 34.3 can be used to evaluate different fab design alternatives to see how (and if) all of the objectives can be met. This section describes key fab design decisions.

34.1.2.1 Resource Choice, Count, and Timing

One of the most fundamental fab design decisions is the number of each type of resource. Examples of resources are process tools and operators. A minimum number of each type of resource will be required to achieve a desired throughput capacity. However, more than that minimum number may be required for some types of tools if the throughput capacity must be achieved within a maximum cycle time constraint or WIP inventory constraint. Section 34.1.1.5 explained that congestion can cause significant queuing when throughput is close to capacity. Thus, extra copies of tools may be added beyond the required minimum in order to reduce queuing in front of those tools. Common candidates for extra tools include relatively inexpensive tools, tool types with potentially long down times that could severely disrupt the flow of material, and tool types where only one tool is required, causing fab performance to be particularly vulnerable to interruptions of that tool.

Factory modeling can also impact the choice of tools. Section 34.1.1.5 also explained that variability, batch processing, and re-entrant flows, all contribute to congestion. For example, long down times can contribute to variability in the fab, so a more reliable tool may be more desirable than a less reliable tool even if the less reliable tool has a slightly lower CoO. Tools that can be purchased in relatively small chunks of capacity may also be desirable (see Section 34.2.4.2). Tool choice criteria that are more difficult to model include technical concerns such as the ability of the tool to perform a process in a repeatable way, or strategic concerns such as the likelihood that the tool would be usable in future process recipes.

Because the demand that a fab is addressing may be growing over time, the targeted fab capacity may also be growing over time. In addition, as yield problems and tool reliability problems are resolved, or as the mix of process recipes in the fab changes, the capacity of an existing fab may also change. All of these factors may motivate the addition of processing resources over time. Timing of resource additions can be particularly challenging because it may be hard to predict demand over time horizons equal to tool lead times or operator training times. Dynamic programming is an example of a method that can be used to optimize the timing of tool additions in such environments [1].

34.1.2.2 Layout

Fabs have historically been laid out so that all of the copies of the same type of tool appear together. There are several motivations for this so-called functional layout. First, process tools have historically been quite unreliable, so it is desirable for all copies of a tool to be grouped together so that the copies can back each other up. Second, there will be some tools that perform multiple steps. For example, 17 steppers may be used to perform 22 masking operations. The steppers are grouped together because it is impossible to have a one-to-one assignment of steppers to masking steps. A third reason is that one operator may be able to run multiple tools (e.g., furnaces), as long as those tools are close enough together that the operator can access any of the tools easily. A fourth reason is that similar tools may share a set of facilities.

Despite all the motivations for functional layouts, there is a growing cost of functional layouts as well. Because functional layouts do not reflect the sequence in which the tools are used in the process recipe, lots may have to traverse long, complicated, overlapping path through the fab as they move from process step to process step. This problem becomes more serious as the number of step increases and as lot transportation is automated. Factory modeling may be used to evaluate different layout alternatives, such as a hybrid between functional layouts and a layout reflecting the process sequence.

34.1.3 Operations Policies

Even for a given fab design, the choice of operations policies can have a significant impact on fab performance. Operations policies are the methods used to make day-to-day operational decisions such as which lot to process next on a tool. In general, operations policies can be grouped into *discrete-review* and *continual-review* policies. Discrete-review policies involve periodically assessing that status of decision criteria (e.g., the lots in a tool's queue), and generating a schedule that specifies the actions to take until the next assessment. Continual-review policies make decisions each time the status of decision criteria changes (e.g., when a lot enters or leaves a tool's queue). For this reason, continual-review policies generally use fairly simple rules, while discrete-review policies may involve complex optimization.

Operations policies are especially used to decrease the effects of congestion (see Section 34.1.1.5) and the resulting queuing and delay. There are entire academic disciplines devoted to studying operations policies. This section describes some of the more common operations policies used in wafer fabs, with an emphasis on continual-review policies. Extensive reviews of operations policies can be found in the literature (e.g., Ref. [21]).

34.1.3.1 Fab Lot Release

Lot release policies are rules that determine when new lots should enter the fab and begin the first process. As described earlier, the random release of lots into the fab can contribute to congestion, increasing the WIP inventory and cycle time for a given throughput. Lot release policies attempt to release lots into the fab in a way that congestion, and therefore WIP inventory and cycle time, can be reduced without reducing long-run average throughput. Release policies can be divided into two categories: *push* policies and *pull* policies. Push policies pre-determine release schedules for lots. For example, a static analysis (see Section 34.2.1) may be used to determine the fab capacity, given the mix of process recipes represented by the lots. A schedule could then be developed to release lots into the fab at a constant rate that is slightly less than the fab capacity. Releasing lots at a constant rate in an order the reflects the desired product mix is the most common type of push policy.

Instead of pre-determining a release schedule, pull policies specify fab states or events in the fab that trigger the release of the next lot into the fab. One common pull release policy is CONWIP, which is an abbreviation of "constantWIP." In this policy, a new lot is released whenever a lot in the fab is completed, in order to maintain a constantWIP inventory level in the fab [18]. CONWIP is a *workload regulation* policy. Workload regulation policies attempt to maintain a constant amount of workload in the fab, where "workload" is defined by the policy. Other than CONWIP, the most common workload regulation policy is maintaining a constant level of work to be performed on the bottleneck resource. Assume, for example, that deep-UV steppers are the bottleneck. In this case, the total number of stepper operations or stepper hours required but not yet performed, summed over all the lots in the fab, would be monitored and as soon as that number fell below some threshold, a new lot would be released into the fab [23]. A more easily implemented variation of this workload regulation policy is to release a new lot into the fab each time the bottleneck performs its final operation on a lot. This policy maintains a constant number of lots that have not undergone that final bottleneck operation. While workload regulation policies generally give better performance than simply releasing lots at a constant rate, they are also generally more difficult to implement, so push policies have continued to be popular.

34.1.3.2 Tool Lot Release

In most cases, process tools will process lots as soon as possible. However, there are some circumstances where tools will wait even when there are lots ready to be processed on the tool. For example, a furnace that can accommodate five lots may only have two lots waiting for a four-hour process. A lot release rule may specify that the furnace should wait for some number of additional lots before beginning the process, in the expectation that processing several lots simultaneously will improve fab performance. More sophisticated versions of these rules will determine whether to begin processing or to wait, based on an assessment of when additional lots in the fab are expected to arrive at the tool [24]. So-called *kanban* rules are another type of rule that can delay processing of lots on idle tools. In general, a kanban is a physical card or some similar mechanism used to trigger the processing or delivery of material. For example, there may be three kanbans for a particular metal etching step. When the etcher begins performing the step on a lot, it will pass the kanban back to the lithography area, requesting more lots to process. The lithography station can only process lots destined for the etching step if there is a kanban at the lithography station. Once such a lot is processed, it travels with the kanban to the etcher. By limiting the number of lots between different pairs of steps in a fab, kanbans keep WIP inventory evenly distributed through the fab and ultimately regulate the flow of lots into the fab. An appeal of kanban rules is that they are relatively easy to implement, even though the resulting fab performance may not be quite as good as the performance resulting from a combination of sophisticated lot release and queue sequencing policies.

34.1.3.3 Queue Sequencing

In typical high-volume fabs, it is common for lots to spend less then 25% of their time in a fab actually being processed. The remaining time is spent in queues waiting for tools or operators to become available. When a resource becomes available, *queue sequencing* policies determine which lot the resource will select to process next. There are many sequencing policies that have been proposed for different types of factories [13], as well as specifically for semiconductor fabs [22]. In wafer fabs, sequencing policies are generally *myopic,* which means that the selection of the next lot to process is based only on the current status of the resource and the queue of lots waiting for the resource. An appeal of myopic rules is their ease of implementation.

The most common myopic rule is first in first out (FIFO), also called first come first served (FCFS). In this rule, the lot that has been waiting the longest gets served next. Because of its ease of implementation, FIFO often serves as the "benchmark" policy in factory modeling studies—a more complicated sequencing policy's value is judged by the extent to which it outperforms FIFO. Another common sequencing rule is shortest process time (SPT), which gives priority to the lot that would require the least amount of process time on the tool. In some situation, SPT can achieve shorter average cycle times than

FIFO. At process tools that perform multiple steps in the process recipe, there may be lots waiting for different processing steps. The last buffer first served (LBFS) rule gives priority to the lot waiting for the processing step that is closest to the end of the process recipe. A related rule is SRPT (shortest remaining process time), which explicitly gives priority to the lot with the least total processing time in its remaining process steps. Last buffer first served and SRPT have also been shown to out perform FIFO in certain scenarios, because they are explicitly oriented toward getting work out of the fab as quickly as possible. When implementing LBFS and SRPT rules, FIFO is typically used to break ties (i.e., among lots waiting for the some process step).

Least slack policies give priority to the lot with the smallest (or most negative) slack. Slack refers to the degree that a lot is ahead of some pre-determined schedule. A lot with less positive slack or more negative slack means that the lot is relatively "late" compared to other lots with more slack. One definition of slack is simply the time remaining until a lot's due date. A potentially more useful definition of slack predicts the date that the lot will finish processing, based on how many more process steps are remaining for the lot. Specifically, when a lot is waiting for step *i*, the lot's slack (s_i) is the difference between the lot's pre-determined due date (δ) and the expected date that a lot will finish processing and exit the fab (ζ_i) given that the lot is currently at step i [8]: $s_i = \delta - \zeta_i$.

Least slack policies differ in the way that they determine the due date and the expected completion date. In a make-to-order environment (e.g., a wafer foundry), the due date may correspond to a completion date promised to a customer. In make-to-stock environment where due dated are not explicitly quoted for each lot, a due date can be calculated as the time the lot was released into the fab, plus some standard offset time that depends on the lot's process recipe. An offset of zero may be used, in which case the slack is always negative. The Fluctuation Smoothing Policy for Variance of lateness (FSVL) rule gives priority to the lot with the most negative slack value, where the due date is set to the release date. Fluctuation smoothing policy for variance of lateness has been shown to reduce the variance of the fab cycle times for lots [12]. The Fluctuation Smoothing Policy for Mean Cycle Time (FSMCT) rule is identical to the FSVL, except that the due date is defined as n/λ for the nth lot to enter the fab, where λ is the long-run average throughput of the fab. Fluctuation smoothing policy for mean cycle time has been shown to reduce the mean fab cycle time. One difficulty in using slack rules is determining the expected remaining time in the fab (ζ_i) for a lot at each step *i.* These values can be based on historical averages from the fab, or from simulations of the fab. In either case however, the value of ζ_i will depend on the sequencing policies. So after the slack policy is implemented for example, the actual values of ζ_i will change, requiring updates to the value of ζ_i used to calculate lot priorities. Those new value of ζ_i will result in changes in the way lots are scheduled, which cause each ζ_i to change yet again. This cycle is repeated until the values ζ_i converge on their final values. Converging on final values is difficult, however, if the mix of process recipes, capacity, long-run start rate, or other features of the fab are also changing. Methods of updating ζ_i to quickly achieve convergence in such environments is still an open research question.

Critical ratio sequencing policies are also commonly used in fabs. The "critical ratio" is calculated as some measure of slack, divided by some measure of remaining work in the fab such as the total number of processing steps or the total remaining hours of processing. For each of these critical ratio policies, the lot with the lowest critical ratio would be processed next.

In any fab, there will be several tool sets that process more than one step in the process recipe. For example, assume an implanter performs two different steps, with a chemical vapor deposition (CVD) operation following the first implant step, and, later in the process recipe, an anneal operation following second implant step. There are several rules used, to assign priority to lots in front of the implanter based on the lot's subsequent process step. For example, the LBFS rule described above would always give priority to lots waiting for the second implant step, with ties broken using a FIFO rule. Other rules assign priorities based on the current size of the queues for one or more subsequent steps. The least work next queue (LOWNQ) gives priority to the last that, after completing the current step, would go on to the subsequent step with the smallest queue. In the implant example, if there were two lots waiting for the CVD step and three lots waiting for the anneal step, then lots waiting for the first implant step would get priority at the implanter. Other sequencing rules use pre-determined target queue levels to assign

priorities. In one such rule, the queue level at the next downstream step for each lot is compared to that next step's target queue level. Priority is given to the lot whose subsequent step has a queue level that is the furthest below its target length. If no queue lengths are currently below their target lengths then the queue length that is closest to its target gets priority. If the target queue levels are two lots for the CVD step and four lots for the anneal step in the previous example, than a lot waiting for the second implant step would get priority at the implanter. Instead of using only the queue levels in front of the subsequent step, the sum of the queue levels for several downstream steps could be compared to target values for those sums. In typical versions of such policies, the queue levels would be summed over all the downstream steps up to the next step on the bottleneck, or alternatively, would be summed over all the downstream steps in the entire process recipe. Typically, the long-run average queue levels are used as the targets in these policies. However, determining the queue targets faces the same challenges as determining remaining lot time in the fab, in the discussion of least slack policies above. Specifically, a change to queue targets will change the way lots are prioritized, which will change the long-run average length of queues in front of each step, leading to an update of queue targets, starting the cycle over again. Furthermore, long-run average queue lengths may not exist if capacity or process mix is always changing. Setting appropriate queue targets is still an open research question.

34.1.3.4 Lot Sizing

Historically, the number of wafers that travel together in a lot has been determined in most fabs by the number of wafers that fit into a cassette. A lot entering the fab would typically consist of 24 or 25 wafers (one cassette), or 48 or 50 wafers (2 cassettes). Wafers can be occasionally scrapped, so lot sizes may be random and smaller for later process steps. In this context, lot sizes are also referred to as "transfer batch sizes" because the lot size is the number of wafers that are transferred together from one step to another. Non-standard lot sizes are occasionally used, however, and lot sizes may get smaller as wafers themselves get larger in the future. Evaluation of the effects of lot size on fab performance is therefore likely to increase.

Motivations for relatively large lot sizes include the following: (1) if wafers are transferred in larger lots, then less material handling capacity is required to move a given number of wafers per period; (2) if wafers are loaded into tools in larger lots, than a given number of wafers can be processed with less time spent on tool and operator overhead activities such as loading lots, logging lots, or pumping lots down in loadlocks. Motivations for relatively small lot sizes include the following: (1) smaller lots will spend less time being processed on single-wafers processing tools, decreasing cycle time; (2) if the status of a tool is determined by measuring completed lots, then less wafers are at risk of being mis-processed before catching a problem with a tool.

34.1.3.5 Routing and Setup

A *routing* policy determines how a lot selects a process tool to perform a given step. The choice of tool matters if the tools are not identical. For example, only one of the candidate tools may be currently set up to perform the lot's process step. In addition, a factory modeler may want to evaluate the impact of restricting the choice of tools that a lot could select from. For example, in some fabs, once the first critical exposure step is performed by a stepper on a lot, then that same stepper must perform all of that lot's subsequent critical exposure steps. Routing is a particular challenge in functional test operations, because tester setups are required not only for different process recipes, but even for different circuits (i.e., mask patterns), and because different probe stations may require different amounts of time to perform the same test sequence on a given wafer.

Setup or change-over policies determine when to perform the next setup on a processing tool. For example, an implanter may be currently set up to perform boron implants, and require a few hours of change-over, conditioning, and testing to switch to arsenic implants. The setup policy would determine when the next change-over should be performed. A common setup policy is "serve to exhaustion." For example, if the implanter was following a serve to exhaustion policy, then it would only switch to arsenic implant when there were no more lots waiting for boron implants. After changing over to arsenic, all lots waiting for arsenic implants would be processed before switching back to boron.

34.2 Static Modeling

Static modeling uses long-run average characteristics of process flows and fabs to determine long-run average performance parameters such as throughput and capacity. In contrast to dynamic modeling, static modeling does not consider detailed dynamics of how lots flow through a fab, so static modeling cannot determine inventory levels or cycle time. As a result, static modeling typically focuses on fab design issues rather than on operational policies. An appeal of static modeling is its ease of implementation—static modeling is typically performed using spreadsheets or relatively simple modeling software.

34.2.1 Capacity and Long-Run Utilization Profiles

34.2.1.1 Description

A utilization profile is a breakdown of how a resource such as a process tool spends its time in the fab. For example, the long-run average utilization of an etcher may be: 10% loading and unloading, 35% processing, 15% down for maintenance or repair, 10% processing test wafers, and 30% idle for lack of lots or operators. Fab resource planning typically uses a spreadsheet to predict average utilization profiles for each type of tool in the fab. The inputs to the model are the long-run average throughput of lots (or wafers) that the fab will have to process, and the amount of time each lot (or wafer) will require of the tool. Additional inputs specify downtime and maintenance information, typically in terms of frequency of failures (e.g., mean time to fail, mean processing time to fail, or mean processed wafers between failures), and repair times (e.g., mean time to repair). For example, there may be one line in the spreadsheet corresponding to ashers. Inputs on that line would specify that there were 25 ashers, that each asher required a total of 14 h of process time per lot (calculated as the time to perform 22 operations requiring 0.5 h per lot and three operations requiring 1 h per lot). The overall fab throughput, located somewhere else on the spread-sheet might be 1.5 lots per hour. Then the average utilization of an asher would be

$$\frac{14 \text{ h/lot} \times 1.5 \text{ lots/h}}{25 \text{ ashers}} = 0.84 \tag{34.3}$$

or 84%. Similar calculations can be performed to determine the fraction of time that the ashers will spend on various other categories of time such as maintenance. If the sum of all of these fraction adds up to a number less than 100%, then there is a sufficient number of ashers to process the specified throughput (1.5 lots per hour), and the ashers will be idle for the remaining fraction of time.

Similar spreadsheet calculations can also be used to determine the required number of machines to sustain some throughput. For example, assume that in the above example, an asher performs an average of 100 operations before failing. There are 25 ash operations per lot, so there is an average of $25/100 = 0.25$ failures per lot. If the average repair time is 2 h, then the average repair time per lot is $2 \times 0.25 = 0.5$ h per lot. As described above, the process time per lot is 14 h, so the total average ash time per lot is 14.5 h. Assume that a 1-h maintenance operation is performed every 10 h, consuming 10% of the asher's time, regardless of how many lots are actually processed, so there are 0.9 h available for processing out of each hour. The maximum long-run sustainable throughput of the asher is thus $0.9/14.5 = 0.0621$ lots per hour. If a fab throughput of 1.5 lots per h is desired, then the minimum number of ashers to sustain that throughput is $1.5/0.0621 = 24.2$, or (rounding up) 25 ashers. Actual spreadsheets may be more complex, including the effects of wafer loss during processing and several other categories of time.

34.2.1.2 Limitations

The long-urn utilization and capacity calculations described above have the advantage of easy, quick implementation, but there are also some limitations of such models. As descried in Section 34.1.1.5, adding extra tools beyond the minimum number required can reduce WIP inventory and cycle time for a given throughput. This effect can not be measured using static models. Also, if static models are used to

calculate the required number of operators as well as of process tools, the static models would neglect the possibility that a tool may be starved for work because no operators are available at that moment to load lots into the bottleneck tool, even though on average each operator will have some idle time. This effect is described as interference between the operator and the tool. Dynamic modeling must be used to determine the extent of interference effects between resources such as tools, operators, and automated material handling systems. Finally, it may be difficult to accurately determine all of the inputs for the spreadsheet model (e.g., time between failures). Some models will compensate for these problems by setting maximum utilizations of less than 100% when calculating the number of required resources.

Static model calculations of the required number of resources to sustain a fab throughput often assume that it is theoretically possible to achieve 0% idle time for at least one tool type if there is a sufficiently high throughput in the fab. In most cases, this assumption is approximately correct, although there are examples of systems with particular queue sequencing rules that can never achieve idle times close to 0% on any of the system's resources [19]. In such a system, even if each tool group had a capacity of 1 lot per hour, releasing 0.9 lots per hour into the system would only result in 0.7 lots per hour (for example) exiting the system, with the remaining lots building up in a continually growing WIP inventory. Dynamic modeling is the only way to check for such unusual conditions in fab.

Finally, there are cases where certain additional assumptions must be made to determine utilization. For example, a furnace may be able to process 6 lots simultaneously, but may only process four lots at a time on average. When calculating the number of furnaces required to sustain a fab throughput, an average number of lots per furnace load needs to be assumed. As another example, a multi-chamber cluster tool may be used to run a sequence of metal deposition steps. Although it may take 1.5 h to process a lot, it may be possible to start processing a second lot while processing on the first lot is being completed. In this case, the cluster would be simultaneously processing two lots for a small fraction of time. In such situations, the lot *takt time* may be used instead of the lot process time when calculating the required number of resources to sustain a throughput. The lot takt time is the average time that passes between the completion of one lot and the completion of a second lot, if the tool is processing at its maximum throughput. The lot takt time is the reciprocal of the tool's maximum lot throughput.

34.2.2 Cell and Cluster Tool Models

An increasing number of processing tools are taking the form of cluster tools, which consist of several processing modules arranged around a central handler, with one or more ports for lots. A lot is loaded into a port, which may then undergo a fixed delay to pump down the pressure in the port for example. Wafers are then moved with a robotic arm (i.e., handler) one at a time to the processing modules. Each wafer may visit one or more module, depending on the cluster configuration. Once all of the wafers are completed and back in the port there may be a final port delay then the lot is unloaded. There may be multiple ports in a cluster tool, allowing several lots to be processed simultaneously.

Static models can be used to estimate the maximum throughput of the cluster tool, as well as the tool cycle time for a lot. The tool cycle time can be closely approximated at $T + lt$, where l is the lot size. If the cluster ports can supply lots to the cluster at a rate that is higher than the rate at which the modules can process the wafers, then the maximum cluster throughput rate (in wafers per time) is $1/t$. The term T is the tool cycle time that must be incurred regardless of the lot size. This term is the sum of the fixed port times (e.g., loading, pumping, etc.) and an initial transient time from when the first wafers begin processing until the cluster operation reaches steady state. This transient time depends on the configuration of the cluster tool, as well as the way the handler is scheduled, and must be measured empirically or determined using dynamic modeling. The term t is the average incremental cycle time resulting from each additional wafer in the lot. This term depends on the number of process modules, the module process times, and the speed of the handler [25].

If the handler speed is not sufficient to supply wafers to the process modules at least as fast as the modules can process the wafers, then the cluster tool's throughput is said to the handler-limited. In this case, reducing module process times will not increase the tool's maximum sustainable throughput. On

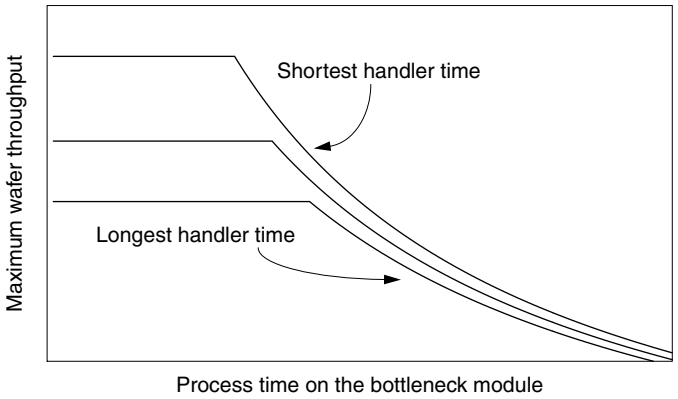


FIGURE 34.3 Maximum wafer throughput of a cluster tool, for three different wafer handling times, as a function of the process time of its most highly utilized module. The throughput is handler-limited in the flat regions on the left and process-limited in the curving region on the right.

the other hand, if the handler's speed is sufficiently high so that it never starves a module for work, then the cluster tool's throughput is said to be process-limited. In this case, reducing either the wafer handling time or the process time of the most highly utilized module will increase the tool's maximum sustainable throughput. Figure 34.3 shows the effect of process times and handling times on the maximum sustainable throughput of a cluster tool. Several papers have derived expressions for T and t, from fundamental parameters such as wafer handling time, module process time, number of wafer handlers, and cluster configuration [14,15,26].

34.2.3 Cost of Ownership

34.2.3.1 Definition and Application

The CoO model is a structured static analysis to compare the costs of different tools or process steps. To facilitate fairness in the comparison, a CoO model will include all costs that can be traced to the tool. The tool cost will typically be expressed as a cost per year, cost over the lifetime of the tool, or cost per processed wafer. "Cost per wafer" is used so that a more expensive tool with a higher throughput capacity can be fairly compared to a less expensive tool with a lower throughput capacity. Cost of Ownership models are typically implemented as spreadsheets, although commercial software is available to perform cost of CoO calculations [4]. Typical inputs to a CoO model include:

- The lifetime of the tool. This is often set to the financial depreciation time (e.g., 5 years).
- Time not available including down time, maintenance time, engineering time, and assist time. Parameters such as down time may be calculated from mean times between fails and mean time to repair.
- The maximum throughput of the tool when it is processing wafers.
- The original cost of the tool, including purchase, transport, and installation.
- The capital cost of the clean room space to accommodate the tool.
- The support cost of the clean room space to accommodate the tool.
- The cost to modify or improve the tool.
- Qualification, labor, and materials costs.
- Labor training costs.
- Utility usage per tool, including electricity and chilled water.
- Consumable purchase and disposal costs. Consumables include gasses, chemicals, and sputter targets.

- Test wafer requirements.
- Maintenance costs, including the cost of spare parts and maintenance contracts.
- Labor costs, including operators, maintenance, and engineering staff. For example, the cost of an operator may be calculated as the fraction of an operator required to operate the tool, multiplied by the salary and support costs to keep an operator on all shifts.
- The cost of lost product wafers. This may be calculated as the product of the accumulated value of a wafer when it reaches the tool's process step, multiplied by the long-run average fraction of product wafers that will be scrapped by the tool.
- The cost of die yield loss. This may be calculated as the product of the value of a wafer that has completed the entire process flow up to wafer probe, multiplied by the long-run fraction of dice that will be lost on product wafers due to defects introduced by the tool (die yield loss per step). In turn, the die yield loss may be calculated as the number of killer defects per area introduced by the step, multiplied by the area of a single die.

Cost of ownership spreadsheets may be several pages long because several parameters may be used to calculate each of the inputs listed above. The CoO model uses the input information to calculate the total number of wafers that will be processed over the lifetime of the tool and the total cost that will be incurred over the lifetime of the tool. The total life-time cost is then divided by the total number of wafers to get the long-run average cost per wafer. Alternative tools or processes can then be evaluated by comparing the cost per wafer on their respective process steps.

34.2.3.2 Model Limitations

A detailed understanding of all input data is required for the successful use of CoO. This is because there are no standard input values. Different companies may use different tool lifetime values, cleanroom costs, wafer values, or virtually any other parameter in the model. By changing these parameters, the relative costs of ownership among alternative tools and processes can be changed. The cost per wafer is particularly sensitive to the assumed total number of wafers that the tool will process over its lifetime. When comparing two alternative tools performing a given process, one tool may have a lower cost than the other tool only at some throughput levels. For this reason, sensitivity analyses of cost on required throughput and other assumed parameters are often performed.

A minor limitation of the CoO model is its neglect of queuing effects. As described in Section 34.1.1.5, highly utilized single tools with long, frequent down times can increase cycle time and WIP inventory. As descried in Section 34.1.1.3, these increases can result in costs that are not caught by the CoO model.

Finally, the CoO model can not be used in a straightforward way to capture interactions between multiple tools. For example, a metrology tool may be used to continually evaluate processing by a previous group of tools. Even if one metrology tool has a higher cost per wafer than an alternative metrology tool, the first metrology tool may be preferable because it more effectively catches processing errors, leading to better process maintenance and die yields of the previous tools. The CoO model does not capture this relative value of the first metrology tool. The CoO model only captures cost information.

34.2.4 Static Financial Projections

34.2.4.1 Overview

Financial projections predict the revenues and cash flows for a fab or group of fabs and assembly and test facilities. These models are usually structured models implemented on spreadsheets. In addition to the cost information used in the CoO model, inputs to financial projections will also include estimates of demand and the revenue ultimately generated by each chip produced in the fab. Financial projections are typically divided into different time periods (e.g., quarters) and different values for input parameters such as demand and die yield are entered for each period. The number of tools in each period may be entered as an input parameter, or it may be generated from the demand and yield parameters. One example of a static financial projection model is Sematech's Cost Resource Model.

34.2.4.2 Scale Economies and Granularity Cost

The purchasing cost of tool sets for differently sized wafer fabs running a typical 0.35μ complementary metal-oxide silicon process is plotted in Figure 34.4. The point furthest to the left on the plot represents the cost and maximum throughput of a fab with one of each type of tool. Successive points are generated by adding one tool to the bottleneck corresponding to the previous point and determining the resulting new cost and throughput. Although the cost for each size of fab would increase with more advanced technologies, the general shape of the plot—a straight line with a positive intercept—is not likely to change. In addition, plots of the facilities cost or the annual fixed recurring fab costs would similarly take the form of a straight line with a positive intercept. A third general feature of such plots is that for very low fab capacities (blow about 4000 wafers per month in Figure 34.4) the fixed costs begin to fall above the straight line. As will be shown in the next paragraph, the positive intercept results in economies of scale for fabs.

Using CoO type of calculations described in Section 34.2.3, one can generate the best achievable cost per wafer for each fab size by calculating the total cost of the entire fab over its life-time and dividing the cost by the maximum number of wafers that the fab could produce over that time. The resulting cost per wafer as a function of fab size is shown in Figure 34.5. In the figure, the contribution of variable costs to the cost per wafer is independent of the size of the fab. The main source of variable costs are the cost of raw wafers and other materials. Fixed costs include both recurring fixed costs (from salaries, maintenance contracts, etc.) and capital costs (from tool purchase and installation, fab construction, etc.). The contribution of fixed costs to cost per wafer decreases as the fab gets larger. To see why, recall that in the previous paragraph, fixed costs were said to be a liner function of capacity, which would take the form: fixed fab costs $= K + \lambda k$ where λ is the maximum throughput (i.e., size) of the fab, K is the positive intercept of the line, and k is the slope of the line. The best achievable cost per wafer is determined by dividing this function by λ, giving cost per wafer $= K/\lambda + k$. As a result of the K/λ term, larger fabs can achieve the smaller costs per wafer shown in Figure 34.5.

Figure 34.5 showed the cost per wafer assuming that each sized fab was producing wafers at its maximum possible rate. This is equivalent to saying that the bottleneck tools in each sized fab are being utilized at 100% of their available time. The utilization of the bottleneck tool is proportional to the fab throughput. Figure 34.6 shows how the cost per wafer depends on the utilization of the bottleneck tool,

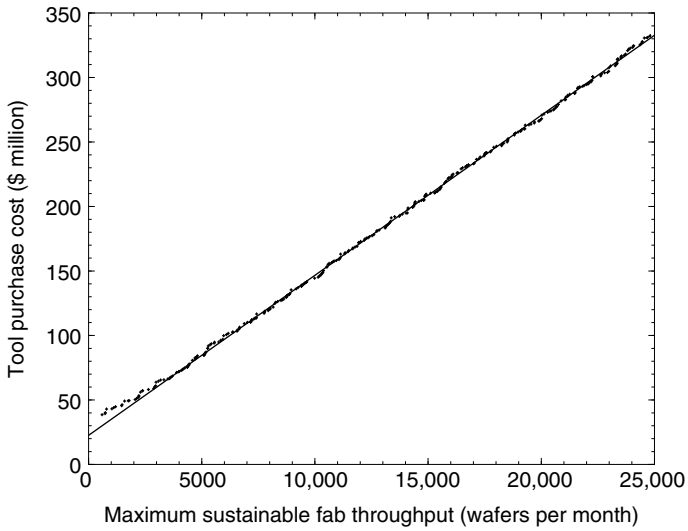


FIGURE 34.4 Each dot corresponds to the tool purchase cost for a differently sized fab. The line was fitted to the dots corresponding to fab sizes greater then 4000 wafers per month.

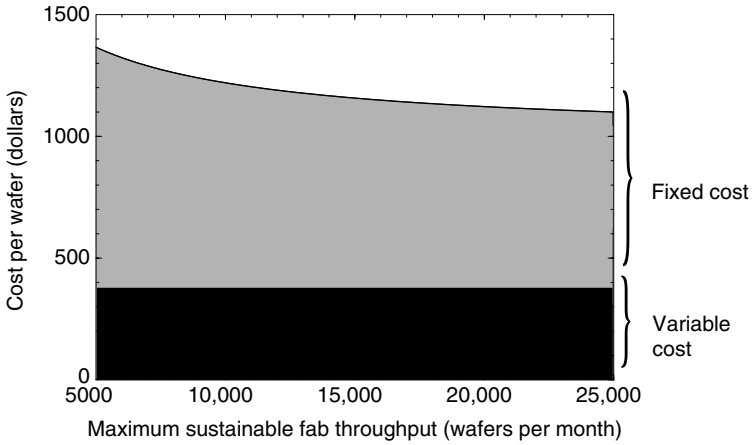


FIGURE 34.5 Cost per wafer in differently sized fabs, assuming each fab is running at its highest achievable throughput and a fab lifetime of 5 years.

for two differently sized fabs. The total cost per wafer is equal to the variable cost per wafer, plus the fixed cost per wafer. The variable costs (e.g., the cost of a raw wafers and photoresist) are proportional to the throughput of the fab, so the variable cost per wafer is approximately independent of bottleneck utilization, or equivalently, independent of fab throughput. The fixed costs (e.g., the purchasing cost of tools or engineering salaries) for a given fab do not change much with fab throughput, so the fixed costs per wafer decrease as the fab throughput (or bottleneck utilization) increases because the same fixed costs are amortized over more wafers.

As described above, determining the effect of fab size on cost per wafer depends on an accurate calculation of K. Costs contributing to K have been called the capacity-independent fixed costs [1]. Granularity costs are an important source of capacity-independent fixed costs. Granularity costs are the cost of extra fractions of non-bottleneck resources that must be purchased but are not required to sustain the fab's maximum throughput. For example, assume a fab is being designed to achieve a maximum

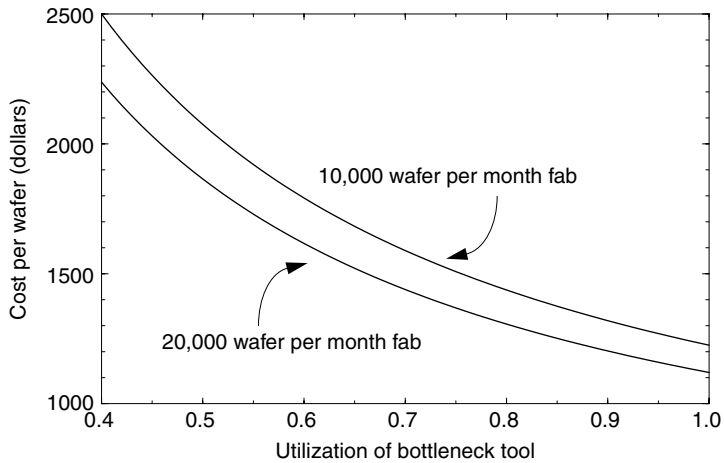


FIGURE 34.6 The cost per wafer for two differently sized fab, as a function of the fraction of available time that the bottleneck tools are processing product wafers.

throughput of 1.5 lots per hour. One of a certain type of tool going into the fab can sustain a long-run throughput of 1 lot per hour, so two copies of that tool are required for the fab. Together, those two tools can achieve a throughput of 2 lots per hour, even though the fab will only produce 1.5 lots per hour. The cost of the extra 0.5 tool that had to be purchased but would not be used has been called the granularity cost. Back in Figure 34.4, the positive intercept is the granularity cost of the tools in the fab. Even though the contribution of each type of tool to overall granularity costs will be different for each size of fab, the total granularity cost of all tools combined is empirically similar for all fab sizes above some threshold (about 4000 wafers per month in Figure 34.4). If tools could be purchased in smaller increments of capacity, there would be less granularity costs, resulting in a smaller value of K, in turn reducing the cost benefits of larger fab sizes.

34.3 Dynamic Modeling

In contrast to static modeling, *dynamic modeling* considers the way that material (e.g., lots) actually queues and moves through a system (e.g., fab). Thus, dynamic modeling can be used to determine WIP inventory levels and cycle times, as well as throughput and capacity. Some of the dynamic modeling techniques described in this section can determine not only long-run average performance value, but also entire distributions of those values. This would be useful, for example, if a fab designer was more interested in the 95th percentile cycle time than the average cycle time. In addition, dynamic modeling can be used to determine performance values during transient periods as well as in steady-state operation. Two general approaches to dynamic modeling are described in this section: simulation (Section 34.3.1) and analytical modeling (Section 34.3.2).

34.3.1 Monte Carlo Discrete-Event Simulation

Simulations of factories are commonly used to predict the performance of a factory with a specific design and set of operations policies. Different simulations can then be used to compare the effects of different fab design decisions or different operations policies. (See Ref. [27] as an example.) Inputs to simulations of water fabs typically include the different process recipes run in the fab. Process recipes are specified as a sequence of steps. Associated with each step are several parameters such as the process time and the required process tool type. The simulation will also have information on each process tool type, including the number of tools of that type, and the wafer capacity of each of the tools. Tool information also typically includes one or more sets of parameters describing the probability distributions of time between failures or maintenance events, and the time to perform repairs and tool re-qualification. The simulation may also include information on operators, configuration of cluster tools, starting lot sizes (particularly if lot sizes can change due to yield loss), and rework probabilities and routes. At least some of the operations policies described in Section 34.1.3 will also be specified in the input data for the simulation. Typical simulation output includes utilization profiles of each type of tool, information on the cycle time distributions of completed lots, average WIP inventory levels, and long-run average fab throughput.

34.3.1.1 Simulation Mechanics and Run Times

A typical *discrete-event simulator* for wafer fabs will track each simulated lot as it moves from tool to tool. An event clock is used to keep track of the simulated time. For example, there may be 100 lots in a simulated wafer fab at simulated time $=200$ h. The next event in the fab is lot No. 55 beginning an oxide etch operation at time $=200.5$ h. The simulation software would update the simulation clock to 200.5 h, and update various data fields for the etcher and lot No. 55. This data would be used, for example, to ultimately calculate the utilization of the etcher or the fraction of time that lot No. 55 spent actually being processed by equipment. If the etch time was 2.3 h, the initiation of the etching event would trigger new events to be scheduled for time $200.5+2.3=202.8$ h. These new events might be lot No. 55 joining the queue in front of the ashers to perform its next step, and the oxide etcher processing the next lot in its queue. After the data for the lot and tool are updated and the triggered future events are scheduled, the

simulator would initiate the next scheduled event which could, for example, be a stepper going down for maintenance at time 200.6 h. Some event durations, such as the time until a tool's next failure, may be random numbers. The name *Monte Carlo simulator* indicates that the simulator includes random number generation. The simulation will include some criteria that triggers its termination, such as the total simulated time or the total number of lots to be completed. Once the simulation is terminated, the detailed data generated by the program is aggregated into useful performance parameters, such as the mean and SD of the lot cycle times and average utilization profiles for each type of tool.

The computer time to run a discrete-event simulation depends on the speed of the computer, the total number of events to be executed in the simulation, and the amount of computer operations associated with each event. For example, the processing of a lot on a cluster tool may be aggregated into one simulated event, or it may be broke down into several component events: loading the lot, transporting each wafer to and from processing modules, processing each wafer in a module, and unloading the lot. Aggregating events cuts simulation time, but may sacrifice simulation fidelity. For example, if the processing of a lot on a cluster tool is aggregated into a single event, then the simulation could not directly determine the effect of faster wafer handling time on tool performance. The number of events to be executed in the simulation also increases as the simulated time increases, or as the size of the simulated fab increases.

34.3.1.2 Probability Distributions and Confidence Intervals

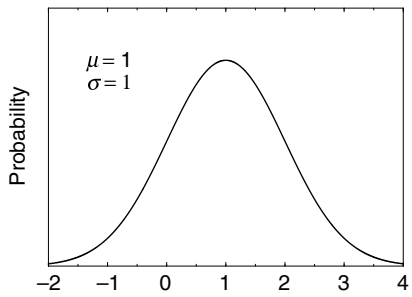
Commercial software for running discrete-event Monte Carlo simulations typically include a choice of several probability distributions for random number generation. The time between a tool's failures and the time to repair a failure are two of the parameters most commonly represented as random numbers. Other manual operations such as lot transportation, loading and unloading of lots, tool setups, and other manual processes may also be specified as having random distributions. Three of the most commonly used random number probability distributions are described in Figure 34.7.

Some simulations may include the possibility of wafer mis-processing, resulting in either rework or wafer loss. As an example or rework, a test following a lithography develop step may detect that some of the wafers were mis-aligned, requiring those wafers to be stripped of their photoresist and go through the previous photolithography sequence again before the entire lot can advance on to the next process step. Mis-processing on other steps, such as implant, may not be correctable through rework, so those mis-processed wafers are simply removed from the lot, changing the lot size. Mis-processing of wafers are typically modeled as Bernoulli trials, which means that each wafer in the lot has the same probability of being mis-processed, and the probability of a given wafer being mis-processed is independent of whether any other wafers were mis-processed. For example, each wafer may have a 0.15% chance of being mis-processed. The simulator would draw a number from a uniform distribution between 0 and 1. If the number drawn is less than 0.0015, then the wafer is considered mis-processed, triggering either a rework or wafer disposal. In reality, the assumption that the probability of different wafers being mis-processed are independent is questionable at best—this assumption is made mainly as a convenience.

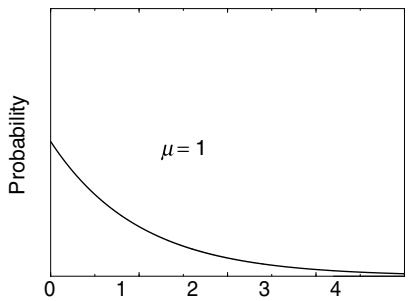
Monte Carlo simulation output is the result of many different random numbers drawn over the course of the simulation, during which the factory behavior is recorded and ultimately aggregated in performance statistics such as average cycle time or throughput. A reasonable concern is whether these statistics would change significantly merely as a result of drawing a different set of random numbers during the simulation. To deal with this concern, performance parameters may be expressed as confidence intervals. To generate a confidence interval, several simulations of the same system will be run, with a different random number generator seed in each simulation so that different random numbers are drawn. The results of each simulation are then recorded and used to generate statistical confidence intervals. For example, if 20 simulations are run, and the average cycle times are normally distributed (even if the cycle time itself is not normally distributed), then the 95% confidence interval for the average cycle time would be

$$\left(\bar{W} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \quad \bar{W} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right) \tag{34.4}$$

Normal Distribution. Variables that are the sum of many other random variables are typically represented using the normal distribution. The distribution parameters are the mean (μ) and standard deviation (σ). Time to repair may be modeled as having a normal distribution. In the rare cases where process times are modeled as random numbers, they are also typically modeled as having a normal distribution. A drawback of using the normal distribution to generate random time durations is that a negative time could be generated. To get around this problem, a simulation program may just throw out negative numbers when they are generated, and then generate another random number.



Exponential Distribution. The exponential distribution is used for its memoryless property. For example, if at any point in time, the time remaining until a tool fails next is statistically independent of the elapsed time since the last failure, then the time between tool failures is said to be memoryless or completely random. This is often assumed to be the case for tool failures, so the time between failures is accordingly modeled as having an exponential distribution. The exponential distribution is also appealing because it is completely specified using only one parameter: its mean (μ) which is equal to its standard deviation. Random numbers drawn from exponential distributions are always positive.



Gamma Distribution. As mentioned before, a normally distributed random variable could have negative values, which has no physical meaning for time durations. To avoid the possibility of negative values, the Gamma distribution may be used. The Gamma distribution may be specified in terms of its shape parameter (α) and scale parameter (β), or it may be specified in terms of its mean (μ) and standard deviation (σ). These two parameter sets are related as follows: $\mu = \alpha\beta$ and $\sigma = \sqrt{\alpha\beta^2}$, which implies $\alpha = \mu^2/\sigma^2$ and $\beta = \sigma^2/\mu$. If $\mu = \sigma$, then the Gamma distribution is the same as the exponential distribution. As μ gets much larger than σ, the Gamma distribution approaches a Normal distribution, except that numbers drawn from a Gamma distribution are never negative.

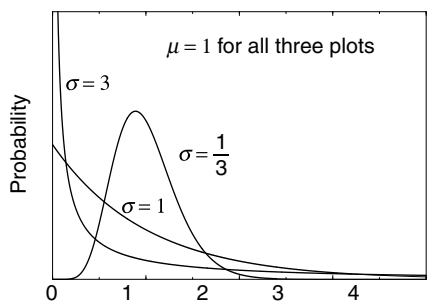


FIGURE 34.7 Three commonly used probability distributions in factory modeling.

where $\bar{W}$ and s are the average and SD of the average cycle times generated by the 20 simulation runs, respectively. The number of simulation runs is represented by n. (In this example $n = 20$.) The term $t_{n-1,1-\alpha/2}$ is the probability that a Student's t distribution with $n-1$ degrees of freedom is less than $1-\alpha/2$. (In this example, $\alpha = 1-0.95 = 0.05$, so $1-\alpha/2 = 0.975$.) Most statistics textbooks or handbooks of mathematical tables will have tables of Student's t distribution. Consulting such a table for this example

yields $t_{19,0.975} = 2.093$. Thus, there is a 95% chance that the true long-run average cycle time is somewhere in the following range: ($\bar{W} - 0.468$s, $\bar{W} + 0.468$s). Statistics books have a similar formula for calculating confidence intervals for the SD of performance parameters. These statistics book would also describe methods (called hypothesis testing) to determine whether performance parameters change in a statistically meaningful way as a result of some change to the simulation input parameters, such as adding a machine or changing an operations policy.

There are a couple of ways of tightening confidence intervals. One way is to increase the number of simulation runs (n). A slightly less obvious approach is to increase the simulated time for each simulation. If the simulated time for each simulation is increased, then the performance parameter values generated by each simulation are more likely to converge on the true performance parameter value. As a result, s will be smaller.

34.3.1.3 Verification and Validation

Input data sets for wafer fab simulations may consist of thousands of parameters. As a result, there is a risk of errors in the input data set. In addition, the simulations software may also have bugs. As a result, modeling using simulation should always include some *verification* that the simulation is performing as intended. There are several approaches to verification. One approach is simply to verify that results are internally consistent. If the steady-state performance of the system is being modeled, then the long-run rate of wafers released into the fab should equal the long-run average rate of completed and scrapped wafers exiting the fab. Furthermore, for steady-state performance, Little's Law should hold (see Section 34.1.1.4). One can also check utilization profile and queue levels for each type of process tool in the simulation. For example, if a process tool is never utilized in the simulation, its corresponding process step may have been omitted from the process flow in the input data set. The utilization profiles of tools can also be compared to the profiles predicted by a static analysis, to uncover problems in the input data or software. If there is a large amount of WIP in the fab, then the fab's throughput should generally be close to the maximum sustainable throughput predicted by a static analysis. Also, by releasing just a single lot into the fab, one can verify that the lot's total cycle time is equal to the minimum cycle time predicted by simply adding up all processing and transportation time. If animation is available, visual checks of the simulation can also be used to verify that it is performing as intended.

Even if the simulation is performing in the way that it was intended, the model may have omitted important features of the fab. As a result the simulation may not accurately model the fab. *Validation* makes sure that the model accurately reflects the fab or system being modeled. Including every minor feature of a fab in a simulation model is probably prohibitively expensive, so it makes sense to focus on the features that will have the greatest impact on the performance metrics of interest. For example, there may be many minor sources of delay for lots moving through the fab, such as operators temporarily waiting for a computer to be freed up to log a lot into the tracking system, or machine down times that have unusual probability distributions. Such effects make it difficult to perfectly model fab cycle times in particular. Even though the model may not perfectly represent the actual fab, it still may be possible to predict the relative benefits of different design alternatives or policy alternatives, as long as key fab features that will be affected by the alternatives are included. Further verification and validation techniques have been surveyed in the literature [2,16].

34.3.2 Queuing Models of Networks

34.3.2.1 Definition and Application

The collection of process tools, lots, and recipes in a fab can be described as a queuing network. If a given tool is busy or otherwise unavailable, the lot enters a queue of lots waiting to be processed by the tool. Queueing models attempt to predict performance metrics such as throughput and cycle time by deriving sets of sophisticated mathematical expressions that describe the flow, and queueing of lots in the fab. Queueing models offer two potential advantages over discrete-event Monte Carlo simulation. First, if the

mathematical equations can be quickly formulated and solved for a given fab configuration, then the queueing model can perform a rough-cut performance analysis much more quickly than a simulation. Second, the equations themselves express the determinants of fab performance, providing insight that can be used to quickly identify the most promising opportunities for fab improvement.

Despite the promise of network queuing models, simulation is by far the more popular approach to dynamic modeling of wafer fabs. There are several features of wafer fabs that make development of accurate queueing models quite difficult. First, process recipes follow re-entrant flows, which means a lot visits a given type of tool more than once in its recipe. Second, the shear number of process steps and tool types results in complicated models. Finally, lots may be processed on either single-lot processors, batch processors, wafer tracks, or cluster tools. This wide variety of process tools is difficult to capture in queueing models.

Because of various complicating features of fabs, queueing modeling of semiconductor fabs still constitutes an emerging science with several different approaches. Two major families of these approaches are descried below.

34.3.2.2 Network Decomposition

One of the most famous queuing results is Kingman's extension of the Pollaczek–Khinchine formula, which approximately describes the queueing of jobs arriving randomly at a single machine [7]. (See also Ref. [6] for a more detailed description.) In the model, jobs are processed on a FCFS basis, and each job has a random process time. The formula is as follows:

$$W_q = M\left(\frac{\rho}{1-\rho}\right)\left(\frac{c_a^2 + c_s^2}{2}\right) \tag{34.5}$$

where W_q is the average time that a job has to wait in the queue, M is the average processing time for a job, ρ is the tool utilization (i.e. the fraction of time that the tool is processing), c_a is the standard deviation of the inter-arrival times between jobs divided by the average inter-arrival time between jobs, and c_s is the standard deviation of the job's process times divided by M. The "inter-arrival time" is the time that passes between two consecutive jobs arriving at the machine or its queue to be processed. If λ is the long-run average rate at which jobs arrive at the tool, then the average inter-arrival time is $1/\lambda$, and the tool utilization is $\rho = M\lambda$. The formula for W_q is plotted in Figure 34.8 as a function of utilization, for different levels of variability. The c_a and c_s terms are measurements of variability. If either of these terms are greater then zero, then as utilization (ρ) approaches 1, queuing time (W_q) approaches infinity.

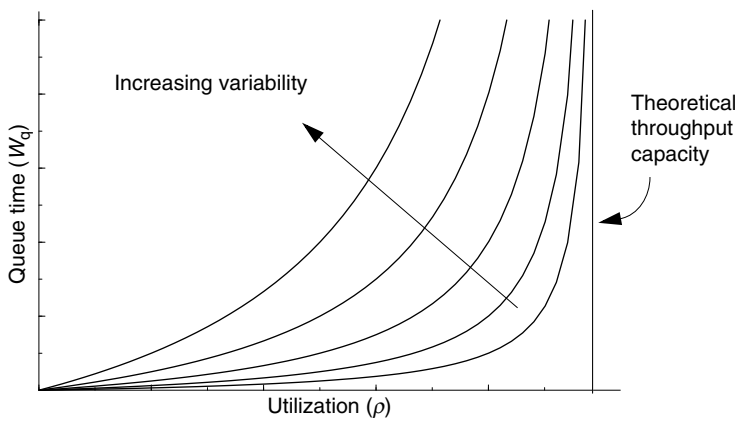


FIGURE 34.8 The modeled relationship between queue time in front of single tool and the utilization of that tool, for different levels of variability in process time and arrivals.

Furthermore, at any utilization level, increasing variability increases queuing. These basic insights generally apply to more complicated queuing systems as well.

A common method of analyzing queuing networks is to decompose the network into individual tool groups and then analyze each tool group in isolation. The analysis of individual tool groups uses equations like the one above. However, the model would ideally be modified to include the possibility of multiple tool copies serving the queue, different queue sequencing rules, different types of tools such as batch tools, cluster tools, or tracks, and tool interruptions such as failures. Queuing theory does not provide exact mathematical implementations of these extensions, but various extensions based on either mathematical derivations or empirical observations can often give reasonably accurate results. A theoretically perfect decomposition process is also unknown, but several approximate techniques exist. Decomposition methods typically determine the throughput of the tool group using a static analysis, estimates the variability of lot arrivals from the characteristics of the other tool groups that feed the tool group of interest, and may estimate the process time variability by including variability from such sources as tool interruptions and operator loading and unloading. Once the decomposition is complete, queue levels are estimated for each tool group and then the total time that lots spend in queues as well as in processing can be determined. See Ref. [3] for an example of a state-of-the-art network queuing analysis of a fab using decomposition methods.

34.3.2.3 Integrated Network Modeling

One shortcoming of using the decomposition methods described above is that only local interaction between process tools are considered. For example, it is very difficult to decompose a queueing system in a way that catches the effect of a scheduling decision at a tool that is visited several times in the process recipe, because the scheduling decision may affect the variability of arrivals of wafers for later process steps at the same tool which in turn could change the effect of the scheduling decision. In this chapter, the term "integrated network modeling" refers to alternative approaches to network modeling that do not use decomposition. These alternative approaches are even less developed than approaches using decomposition, and involve considerable mathematical sophistication.

One recently developed integrated network modeling approach [9] begins by identifying theoretical constraints on certain features of the queueing network, such as the throughput capacity of each tool, or some function of the average queue length in front of the tool. Some of these constraints may be different for different queue sequencing policies, for example. Once a set of constraints are generated, other mathematical techniques such a linear programming are used to find the minimum and maximum values of a selected performance parameter (typically cycle time) that do not violate the constraints. In this way, one can determine upper and lower bounds on fab performance parameters that result from different scheduling policies or tool counts.

Another broad approach to integrated network modeling (see [5] for example) begins by scaling the movement of material moving through the fab as a continual flow of infinitely divisible work moving through queues and tools. This scaling allows the flow and queueing of material to be expressed as equations with continuous variables. The equations are then used to estimate performance parameters (typically cycle time) of the scaled network, providing insights or performance parameter values for the original network. There is more than one mathematical type of scaling the original network. One type, often called a *fluid approximation,* models lots as a fluid, and tools as pumps. Another type of scaling assumes a different type of continual motion of material, called Brownian motion, with special mathematical properties that make it easier to capture the effects of variability, but also requires more mathematical sophistication.

Acknowledgments

The author thanks Sunil Kumar for stimulating discussions of this chapter's material.

References

1. Angelus, A., E. L. Porteus, and S. C. Wood. Optimal Sizing and Timing of Capacity Expansions with Implications for Modular Semiconductor Wafer Fabs. Stanford University Graduate School of Business Research Paper No. 1479. December 1997.
2. Balci, O. "Validation, Verification, and Testing Technique throughout the Life Cycle of a Simulation Study." In *Proceeding of the 1994 Winter Simulation Conference*, edited by J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 215–20. Orlando, FL: ACM Press, 1994.
3. Connors, D. P., G. E. Feigin, and D. D. Yao. "A Queuing Network Model for Semiconductor Manufacturing." *IEEE Trans. Semicond. Manufact.* 9, no. 3 (1995): 412–27.
4. Cost of ownership for semiconductor manufacturing equipment metrics. Standard E-35. Semiconductor Equipment and Materials International, Mountain View, California.
5. Dai, G., D. H. Yeh, and C. Zhou. "The QNET Method for Re-entrant Queuing Networks and Priority Disciplines." *Operat. Res.* 45 (1997): 610–23.
6. Hopp, W. J., and M. L. Spearman., *Factory Physics*. Chicago, IL: Irwin, 1996, chap. 8.
7. Kingman, J. F. C. "The Single Sever Queue in Heavy Traffic." *Proc. Cambridge Phil. Soc.* 57 (1961): 902–4.
8. Kumar, P. R. "Scheduling Semiconductor Manufacturing Plants." *IEEE Control Syst. Mag.* 14, no. 6 (1994): 33–40.
9. Kumar, S., and P. R. Kumar. "Performance Bounds for Queuing Networks and Scheduling Policies." *IEEE Trans. Automatic Control* 39, no. 8 (1994): 1600–11.
10. Leachman, R. C., and D. A. Hodges. "Benchmarking Semiconductor Manufacturing." *IEEE Trans. Semicond. Manufact.* 9, no. 2 (1996): 158–96.
11. Little, J. D. C. "A Proof for the Queuing Formula: $L = \lambda W$." *Operat. Res.* 9, no. 3 (1961): 383–7.
12. Lu, S. C. H., D. Ramaswamy, and P. R. Kumar. "Efficient Policies to Reduce Mean and Variance of Cycle-Time in Semiconductor Manufacturing Plants." *IEEE Trans. Semicond. Manufact.* 7, no. 3 (1994): 374–88.
13. Panwalker, S. S., and W. Iskander. "A Survey of Scheduling Rules." *Operat. Res.* 25, no. 1 (1977): 45–61.
14. Perkinson, T. L., P. K. McLarty, R. S. Gyurcsik, and R. K. Cavin. "Single-Wafer Cluster Tool Performance: An Analysis of Throughput." *IEEE Trans. Semicond. Manufact.* 7, no. 3 (1994): 369–73.
15. Perkinson, T. L., R. S. Gyurcsik, and P. K. McLarty. "Single-Wafer Cluster Tool Performance: An Analysis of the Effects of Redundant Chambers and Revisitation Sequences on Throughput." *IEEE Trans. Semicond. Manufact.* 9, no. 3 (1996): 384–400.
16. Sargent, R. G. "Verification and Validation of Simulation Models." In *Proceedings of the 1994 Winter Stimulation Conference*, edited by J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 77–87. Piscataway, NJ: IEEE, 1994.
17. Guideline for definition and measurement of equipment reliability, availability, and maintainability (RAM). Standard E-10. Semiconductor Equipment and Materials International, Mountain View, California.
18. Spearman, M. L., D. L. Woodruff, and W. J. Hopp. "CONWIP: A Pull Alternative to Kanban." *Intl J. Prod. Res.* 28, no. 5 (1990): 879–94.
19. Seidman, T. I. "First Come, First Served Is Unstable!." *IEEE Trans. Automatic Control* 39 (1994): 2155–71.
20. Standard for definition and measurement of equipment productivity. Standard E-79. Semiconductor Equipment and Materials International, Mountain View, California.
21. Uzosy, R., C-Y. Lee, and L. A. Martin-Vega. "A Review of Production Planning and Scheduling Models in the Semiconductor Industry, Part I: System Characteristics, Performance Evaluation, and Production Planning." *IEEE Trans.* 24, no. 4 (1992): 47–60.
22. Uzsoy, R., C-Y. Lee, and L. A. Martin-Vega. "A Review of Production Planning and Scheduling Models in the Semiconductor Industry, Part II: Shop-Floor Control." *IEEE Trans.* 26, no. 5 (1994): 44–55.

23. Wein, L. M. "Scheduling Semiconductor Wafer Fabrication." *IEEE Trans. Semicond. Manufact.* 1, no. 3 (1998): 115–29.
24. Weng, W. W., and R. C. Leachman. "An Improved Methodology for Real-Time Production Decisions at Batch-Process Work Stations." *IEEE Trans. Semicond. Manufact.* 6, no. 3 (1993): 219–25.
25. Wood, S. C., S. Tripathi, and F. Moghadam. "Generic Model for Cluster Tool Throughput Time and Capacity." *Proc. IEEE/SEMI Adv. Semicond. Manufact. Conf.* (1994): 194–9.
26. Wood, S. C. "Simple Performance Models for Integrated Processing Tools." *IEEE Trans. Semicond. Manufact.* 9, no. 3 (1996): 488–94.
27. Wood, S. C. "Cost and Cycle Time Performance of Fabs Based on Integrated Single-Wafer Processing." *IEEE Trans. Semicond. Manufact.* 10, no. 1 (1997): 98–111.

35

Economics of Semiconductor Manufacturing

35.1 Introduction **35**-1
35.2 Market and Manufacturing Dynamics **35**-1
35.3 Moore's Law **35**-2
35.4 Economic Effects and How Manufacturing Fits In **35**-4
Rapid Growth Cycles • Capacity and Market Pricing
35.5 Economic Models **35**-9
35.6 The Learning Curve **35**-9
35.7 The Technology Treadmill **35**-12
35.8 Technological Driving Forces **35**-14
Technological Pace • Evolutionary Technical Change • Revolutionary Technical Change • Characteristics of Technological Driving Forces
35.9 Factory and Equipment Economics **35**-18

G. Dan Hutcheson
VLSI Research, Inc.

35.1 Introduction

The semiconductor story is unparalleled in the history of mankind largely because of its effect on accelerating the advance in technology and the resulting effect on the world's economy in the second half of the 20th century. When the semiconductor was born into the world at Bell Labs in 1948, its announcement was largely unheralded. The New York Times relegated it to a back page. It would not really begin its relentless march to international fame until 1959; when projections of the computing needs to send a man to the moon indicated the computer would have to be sized as large as the Empire State Building and would consume enough power to light up the eastern seaboard of the United States. At that time, a megabit of memory cost $75 M. Today, it is around 15 cents. This chapter examines how technology, manufacturing, and economics have intermingled to make this stunning history possible.

35.2 Market and Manufacturing Dynamics

Semiconductors have been at the center of high technology for over half a century. Few other scientific topics have drawn so much of the world's attention. Semiconductors have become of such importance that many governments around the world view them as a strategic resource. However, in the greater scheme of things, semiconductors are part of an intense history of demand for information storage and retrieval that dates back to Early Sumerian clay tokens, winds its way through the Gutenberg printing

press, the Dewey decimal system, and eventually results in the semiconductor. Throughout history, the increasing ability to store and process information has been an important driver of legal, social, economic, scientific, and business processes. Moreover, today's semiconductor industry is the foundation of a trillion-dollar electronics industry that provides tens of millions of jobs. It is this foundation that, during the 1990s, drove high technology to become the leading source of economic and job growth in the industrialized world. Surprisingly, these dividends are largely due to the semiconductor industry's ability to consistently double the number of transistors that can be fit on a chip for the same cost.

35.3 Moore's Law

In 1965, Dr Gordon Moore observed that the number of transistors that on a chip was doubling every year as advances in lithography shrunk critical dimensions (CDs). Moore, who is a co-founder of both Fairchild and Intel Corporation, had created what he would later call a self-fulfilling prophecy. The phenomenon became known as Moore's Law, and it has had far-reaching implications for high technology and society because the doubling in density was not accompanied by an equivalent increase in cost. Thus, the average cost of a transistor was significantly cut. Higher levels of integration also meant that greater functionality could be integrated onto a chip. Making the device elements smaller also meant that the transistor's device became faster. The resulting increase in computing power for the same price meant that more complex software could be developed, which pushed the demand for even more power. This feedback loop resulted in increasing sales of chips, computers, and software, leading to the information revolution that has transformed society today.

The principle applies to all kinds of chips and has been remarkably consistent for an industry subject to such strong cycles, as shown in Figure 35.1. From just a few thousand transistors in the early 1970s to today's 100 M transistor chips, the curve is almost straight. Even more surprising is its acceleration since 1995. The reduction in cost is just amazing. Let us use memory as an example, because it is the highest volume business and easiest to measure the economic consequences. From about $13,000, in 1971, the price of a megabyte of semiconductor memory has declined systematically, due to Moore's Law, breaking orders-of-magnitude barriers every few years. Economists like to measure value in what they call "real" terms, which means without inflation. The deflation of prices in the semiconductor industry has had an almost unimaginable effect on the world's economy. To put this in perspective, the real value of the world's semiconductor production, in constant transistor prices, now orders-of-magnitude greater than

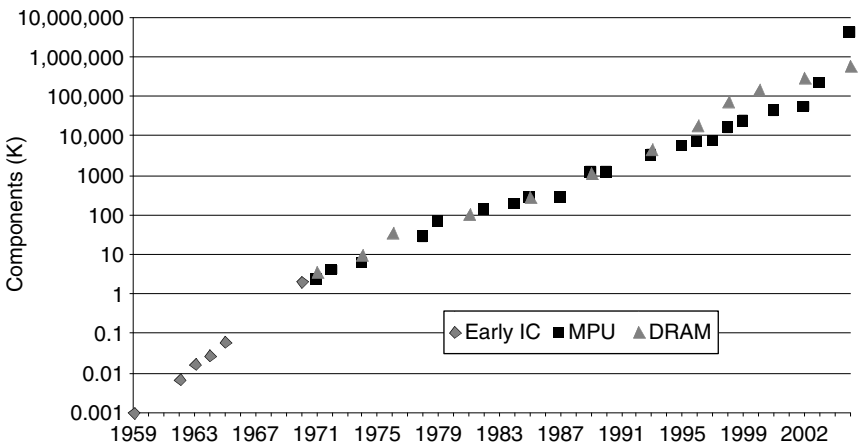


FIGURE 35.1 Moore's law.

the value of Gross world product. At no time in history has any technology delivered such an economic benefit.

On the surface, it may seem that Moore's law left a clear and open roadmap to the dependable doubling of transistor densities. However, the roads have proven harrowing with no clear path. There have been repeated barriers in equipment and production processes that have had to be overcome. The roots of Moore's law lie in manufacturing, for its advances in manufacturing that have made Moore's law possible.

The economics behind these breakthroughs are shown in Figure 35.2. With any given manufacturing technology, the costs of achieving higher levels of chip performance rise very rapidly as its limits are approached and then surpassed. Increasing costs eventually drive prices beyond what buyers are willing to pay, causing the market to stagnate before the physical barriers of technology are encountered. At this point, the industry must jump from the cost-performance curve associated with an old technology to a new technology's cost curve. Switching to a new manufacturing technology renews the downward drive fabrication costs. In effect, the breakthrough from one manufacturing technology to another forces the cost curve to bend down and outward, pushing technical limits farther out. When this happens, higher levels of performance are obtainable without an increase in cost, prompting buyers to replace older equipment. This is important in the electronics industry, because products seldom wear out before becoming obsolete.

The cost of overcoming these barriers has been high as measured by the cost of building and equipping a new wafer fab. The cost for high volume state-of-the-art wafer fabs memory chips rose from less than $4 million in 1970 to well over a billion dollars (see Figure 35.3). Chip fabrication plants are among the most expensive in the world. While Moore's law was doubling transistors every 18 months during the 1980s, fab costs were increasing by 1.7X in the same period. It is amazing that the industry was able to survive, while trying to hold down the costs of the chips themselves. In the nineties, Moore's law increased back to a doubling every year. Fortunately, the rate of growth in fab costs slowed to a compounded rate or 14%

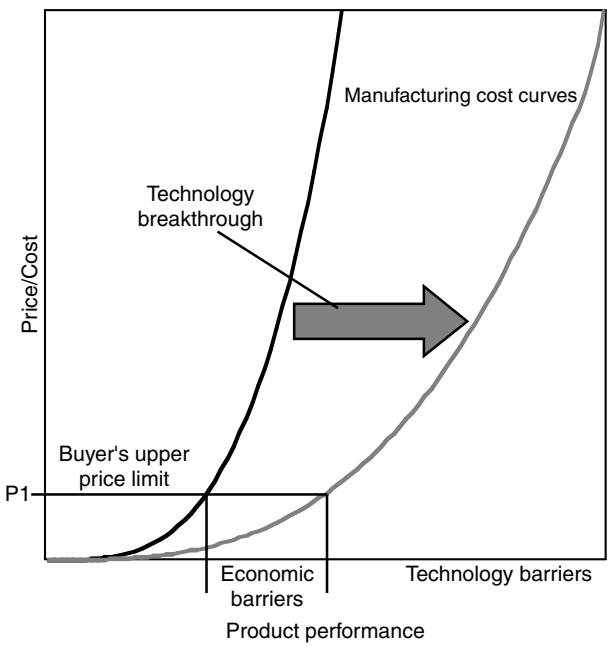


FIGURE 35.2 Effects of technology breakthroughs.

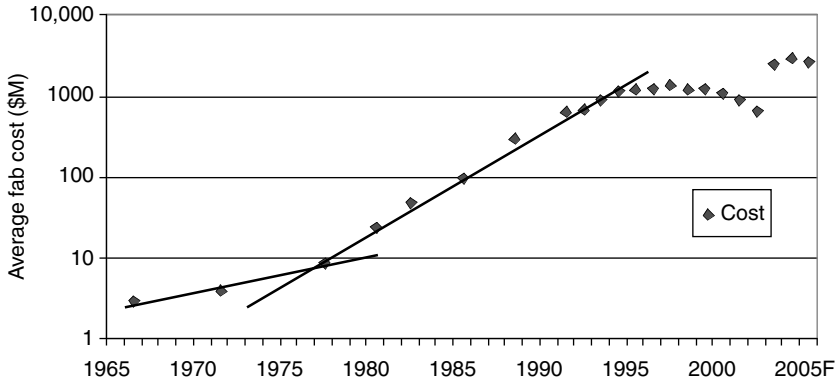


FIGURE 35.3 Trends in the wafer fab costs.

each year, or the industry would be facing $10B fabs today. At that time, even today's cost of $1.5B has put the business beyond the reach of all, but the largest of firms. Such skyrocketing costs, propelled mainly by the expense of having to achieve ever more imposing technical breakthroughs, continue to focus attention on limits in the semiconductor industry. It has been impossible to predict exactly when this stream of improvements that drive Moore's law will dry up. Nevertheless, as the stream becomes a trickle, the economic consequences of approaching technical barriers will be felt before the barriers themselves are reached.

35.4 Economic Effects and How Manufacturing Fits In

The effect of semiconductor technology on the economy is immensely complicated because so much of it occurs in a free market not subject to controls. Both demand and supply play important roles in the market. Consequently, manufacturing and technology dramatically affect the semiconductor market. Understanding the characteristics of the semiconductor market is critical in any decision making regarding manufacturing.

The semiconductor market is well known for its market instability. An understanding of how the dynamics of the semiconductor market drive its instability helps in identifying the early warning signals of an impending shift in market conditions.

There are several key market features that play heavily in causing the industry's market instability. The four most important ones are listed below:

- Rapid growth cycles
- Sharp price fluctuations
- Rapid technical innovation
- Frequent capacity imbalances.

These factors have a strong effect on the ability of companies to compete effectively. Boom-bust cycles are characterized by the entry and exit of firms into the marketplace. For example, in the mid-1980s, there were a multitude of new entrants into the Dynamic Memory Market–Nihon Minebea Co. Ltd. (NMB), Micron Technology, Hyundai, and Samsung to name a few. There were also many who exited the market during the subsequent downturn. Advanced micro devices (AMD), Intel, National Semiconductor, Mostek, and Motorola were among the most prominent. By the late 1990s, Micron and Samsung had

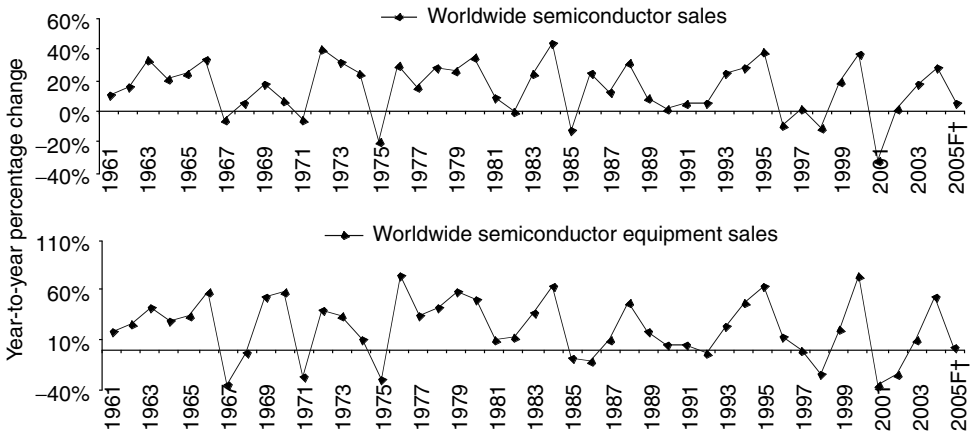


FIGURE 35.4 Semiconductor business cycles.

become giants, while many leaders from the mid-1980s were on the verge of collapse and NMB had failed. Matching a manufacturing strategy with a marketing strategy while maintaining the technological pace is difficult for even the best companies. However, the penalty for failure is quite severe.

These factors also have a dramatic effect on equipment companies. There is an old story that has circulated the industry for years. It goes like this: "when the economy gets a chill, electronics catches a cold, semiconductors come down with pneumonia, and the equipment industry dies." This is an apt description of how the market dynamics affect equipment suppliers.

This section examines each source of instability and their affect on the semiconductor manufacturing industry.

35.4.1 Rapid Growth Cycles

One of the most documented characteristics of the semiconductor market has been its boom-bust cycles. In boom times, the semiconductor market has achieved growth rates that have exceeded 40%. In downturns, sales have fallen by as much as 20% (see Figure 35.4). The effect of these cycles on the semiconductor equipment market is amplified by the acceleration principle.[1] Sales gyrate sharply with small changes in semiconductor growth. Equipment bookings have risen in excess of 100% in an upturn and fallen by over 95% in downturns. Managing a business under these conditions is difficult at best.

The market dynamics for how the various factors interplay to cause severe boom and bust cycles in the industry are explained below:

Production in the electronics industry picks up rapidly once an upturn is underway. This causes an increase in semiconductor orders. Soon, the semiconductor industry is also well on its way to recovery. When this happens, inventories maintained by users decreases to very low levels, drawing up demand still further. The semiconductor industry soon becomes capacity-limited. Shipments slow while delivery lead times stretch out. Longer delivery lead times require that electronics manufacturers keep still more inventory in order to keep their production pipelines full; further aggravating the whiplash. This effect creates a strong feedback loop that amplifies semiconductor market cycles.

Electronics manufacturers must keep a sufficient supply of semiconductors on hand to support their production needs. The supply of inventory that is needed has a direct cause–effect relationship on lead times. As lead times expand, still more inventory is needed to account for the longer delivery times.

[1]Keiser, N. F. Macroeconomics, Random House, 157–158, 1975.

Consequently, electronics manufacturers increase purchasing activities. However, these additional purchase orders are not usually filled immediately because of semiconductor capacity limitations. This causes lead times to stretch out even further, which in turn causes still more purchasing activity, while inventories grow to disproportionate levels. This fundamentally unstable situation feeds upon itself causing semiconductor sales to expand without limit.

Invariably, this supply shortfall causes an immediate strengthening of prices. Companies already engaged in the market become immensely profitable as they ride down the learning curve. This excess profitability creates demand for more capacity. Moreover, it attracts new competitors. Meanwhile, semiconductor bookings, lead times, and inventories continue to rise. These effects translate, in turn, into a literal explosion of sales of equipment to manufacture the semiconductors. Demand for such equipment exceeds supply and so equipment backlogs also grow dramatically. Additionally, equipment manufacturers are somewhat reluctant to expand capacity at a rate fast enough to satisfy the industry. Consequently, equipment backlogs usually swell to unprecedented new highs in each new upturn.

But new equipment that has just been installed does not usually become operational for about 9 months. It is new, it is complex, and it is difficult to learn to operate correctly. Moreover, delivery of ordered equipment often takes another 9–15 months. Consequently, the overall lag between an initial market upturn and added online capacity will typically exceed 2 years. By this time, the initial cause of the upturn may have faded.

Two dramatic time-displacement effects take place as a result: first, semiconductor manufacturers, observing this demand upon their already-limited capacity come under great pressure to buy still more equipment, for they are observing the effect of equipment purchases which occurred some 2 years earlier. Their objective will be to get in equipment as quickly as possible.

Secondly, at the same time, the supplier will be booking the equipment planned for delivery 1–2 years in the future. So the suppliers' vantage point is 2 years into the future. This creates a perceived discrepancy of about 4 years between that of the seller and that of the buyer. This displacement between booking, delivery, and increased capacity causes the buildup of a pressure front between supply and demand. Anticipatory effects result and they often come to dominate.

Eventually, three events occur:

- Semiconductor equipment manufacturer's bookings catch up with real demand levels.
- Electronic demand softens.
- New semiconductor capacity comes online to satisfy order levels.

Excess orders cause equipment manufacturers to build equipment at significantly higher rates. Semiconductor manufacturers soon caught up to the real level of electronics demand (i.e., that level of demand that does not include excess inventory, orders, or backlogs).

Semiconductor shipments expand rapidly as new capacity comes online. Growth quickly exceeds demand. In 1995, semiconductor sales grew by 38%. One tier up, overall electronics sales grew by 9%. This was much higher than needed. The capacity ramp, started in 1993, began to come online in the middle of the year. Component lead times started to fall and inventory requirements fell also. The overall effect reversed and anticipatory effects again came to dominate as the industry stalled and then fell. When electronic demand softens, it will often trigger an industry downturn. The effect that these cycles have upon prices is dramatic, as will be seen in the following text.

35.4.2 Capacity and Market Pricing

There has been a long history of rapid price movements in semiconductors, as shown in Figure 35.5. These historical price fluctuations are the most visible cause of the semiconductor cycle. However, there is a complex interaction of underlying factors drive these fluctuations.

Average semiconductor prices have tended to rise over the long term. At the same time, prices for any individual integrated circuit (IC) have typically fallen over the long term. This dichotomy occurs because the rate at which new higher priced ICs are brought to market and the market growth for these ICs are

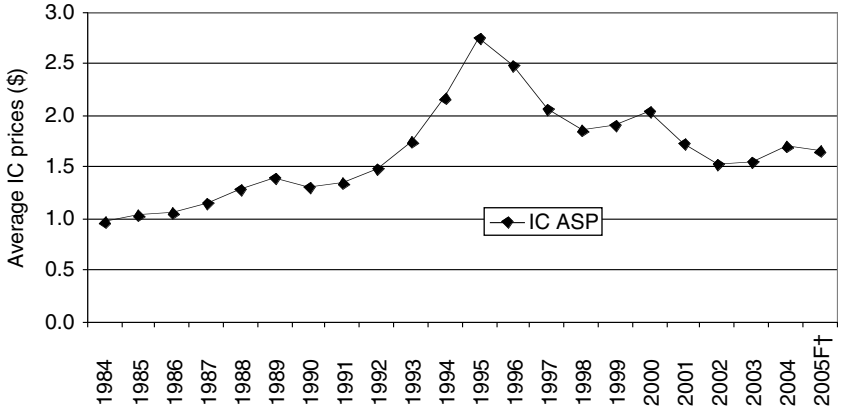


FIGURE 35.5 Metal-oxide semiconductor (MOS) IC price history (Average in $).

greater than that of mature devices. In order to grow, a chip manufacturer must constantly introduce new designs with high market value to offset price declines for older designs whose value is sinking as they become commodities. It is much like the fashion industry, where today's highly priced fashions quickly become tomorrow's discount store bargains, as competitors figure out how to copy the market leader.

When a new product is introduced, prices are usually quite high. A range of $70 to as much as $1000 is typical depending on supply conditions. The individual price of a specific product will usually drop quite rapidly following its introduction-usually down to about 10–$20. As the component matures, the price will settle further, down into the $5–$10 range. Upon reaching maturity, it becomes a commodity item, and will typically cost between 2 and $4. Its price can reach as little as $1 or below. As the device matures, there is often a collapse in prices. This is particularly true in dynamic random access memories (DRAMs) because of the historical capacity swings.

Capacity shortages can also have the opposite effect and can bring rapid price rises. This effect was exhibited most recently in 1993–1995 when a combination of low semiconductor inventories and a sudden increase in computer sales created a surge in semiconductor demand. Several years of low capital investments had left the semiconductor industry without enough capacity. The average selling price for all ICs rose by 13% in 1994 and by another 12% in 1995. Figure 35.6 shows capacity utilization for the

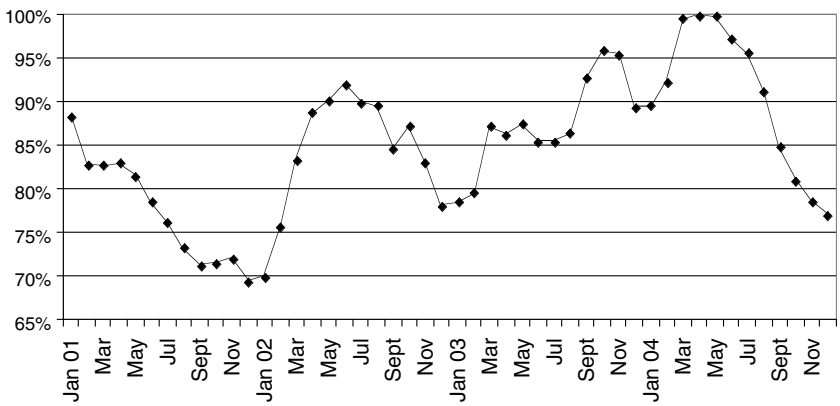


FIGURE 35.6 Capacity utilization.

semiconductor industry as it made the transition from shortage conditions to a glut between 1995 and 1996. The effect of this transition was readily translated into prices in 1996, which fell by 16%. The capacity swing was worst in DRAMs, for which prices rose by 39% in 1994 and by another 35% in 1995. But there were almost a hundred new DRAM fab construction projects announced between 1993 and 1995. The first ones began to come online in 1995. A capacity glut emerged late in the year. Prices first weakened and then, by January on 1996, they were falling. When 1996 had ended, the average price decline recorded for the year in DRAMs was 45%.

One might think that these price variations are purely market related and thus random. However, they have had a periodicity of approximately 5 years. Periods of price collapse have begun in 1974, 1981, 1985, 1991, 1996, and 2001. Shortages have begun in 1973, 1977, 1984, 1987, 1993, 1999, and 2004. Market conditions do play an important role. However, lead–lag relationships in capacity building play a more significant role in creating disequilibrium[2] in these markets.

The history of price collapse in the semiconductor industry indicates that *short-run* demand for semiconductors is highly inelastic.[3] The sharp fluctuations in price, over 6-month period, are strong evidence of this. Such inelasticity occurs because price drops are not readily translated into new demand for devices. Many electronics markets are not very price competitive and those that are tend to be at rock bottom prices. So cost reductions are not immediately translated into lower equipment prices. Moreover, it takes time to design semiconductors into new applications, which could bring about demand for greater volume. The result is that demand for semiconductors does not respond quickly to price changes. This aggravates price fluctuations and makes the market price inelastic in the short run.

This short-run inelasticity was demonstrated in a study of the 64 K DRAM market during from 1984 to 1985 period.[4] As prices fell in 1985, unit volumes did not skyrocket as might be expected in an elastic market. Rather, they stayed reasonably constant. For every additional 89 M units produced in 1985, prices dropped by one dollar, clearly indicating an inelastic market. The specific short-run price elasticity of demand for 64 K DRAMs was measured at 0.083 in 1985. The impact of this was that every additional fab module brought online in 1985 reduced the absolute size of the DRAM market by $545 M!

Because prices are inelastic, the opposite can occur when demand increases sharply after a period of under-investment in capacity. This happened in late 1987 and prices of 256 Kb DRAMs skyrocketed. By mid-1988, prices were almost four times late 1987 prices. Yet, unit volumes were *increasing*. It happened again between 1993 and 1995 with 1 and 4 Mb DRAMS. As a result, excess profitability ensued, attracting too much investment, which lead to a price collapse in 1996.

As can be seen, price instability in the chip market is very sensitive to capacity. It is also very sensitive to technical innovation. Semiconductor prices tend to be capacity-driven in the short run and technology-driven in the long run.

---

[2]Disequilibrium is a term used by economists to describe an imbalance between supply and demand. Disequilibrium in a market results in unstable prices. The pricing history of DRAMs indicates the semiconductor market that is highly elastic in the *long run* over several years. The continuous decrease in price per bit over time and the subsequent explosion in memory demand are proof of this. In fact, the specific *long-run* price elasticity demand for DRAM bits is 6598, as based on measurements between 1977 and 1984. Hence, demand increased by almost seven thousand bits for every one-cent drop in price per Kilobit during this period. An elasticity number greater than one is considered elastic. Consequently, *long-run* price elasticity of demand for DRAMs is extremely elastic. While the author has not measured price elasticity for the overall market, the fact that transistor consumption by the world has grown many orders of magnitude as Moore's Law has brought prices down is ample evidence that the overall market is elastic in the long run as well.

[3]Because prices are inelastic, the opposite can occur when demand increases sharply after a period of under-investment in capacity. This happened in late 1987 and prices of 256 Kb DRAMs skyrocketed. By mid-1988, prices were almost four times late 1987 prices. Yet, unit volumes were *increasing*. It happened again between 1993 and 1995 with 1 and 4 Mb DRAMS. As a result, excess profitability ensued, attracting that the industry may have learned this and has adapted in the post-Y2K environment to controlling capacity. If this proves to be the case, one can expect less cyclicality than in the past.

[4]Hutcheson, G. D., and J. D. Hutcheson. *The VLSI Capital Equipment Outlook*. Vol. 1. VLSI Research Inc., 1985.

35.5 Economic Models

The economics of the semiconductor industry are largely driven by its interaction with technology. Technical innovation affects semiconductor market dynamics in two ways. The first is through new device technology. The second is through the learning curve.

The onset of a new generation of device technology can be a triggering mechanism that initiates a price collapse. Demand for an old generation falls off rapidly when a new generation becomes available.

Often, semiconductor manufacturers will find themselves in a crushing vice of economic forces. Their new capacity for current generation parts is finally coming online. However, the next generation is already in pilot production. Customers are anxiously awaiting the new generation part. Once a word of its imminent production gets out, demand for the current generation quickly falls off. Sagging demand and new capacity causes a price collapse. Losses will begin to steadily mount during this time. This is the most critical point of time, for a manufacturer must expand capacity for the new generation part. Its existing production facilities are obsolete. The last thing that a company wants to do, under these conditions, is invest more in DRAMs. Those that succumb to this urge and do not make the investments, eventually fail. Those that do will have the capacity to reap extremely profitable years in the next capacity shortage.

New technology and failure to invest creates a technology shortage. For example, it only cost about $8.00 to manufacture and sell a 16 Mb DRAM in 1995. At an average price of $14, the net pretax profits were $6 per device. This is why so many companies find the DRAM market so attractive. This is especially true for new market entrants who have not experienced the downside risk of a price collapse. Until recently, there has always been a group of new market entrants anxiously awaiting the next upturn.

While it is too early to say what the effects are, the tectonic changes that occurred after the Y2K debacle have dramatically changed the industry. The are fewer entrants into the industry, as a result of the fact that there have been few investors willing to participate in start-ups. China's efforts have proven to be more press releases than real serious competition. Certainly the history of seeing 10, 20, or more start-ups in an upturn truly seems to be a thing of the past. At the same time, the technology has become extremely difficult to copy. Also, the economy of scale for an efficient factory has doubled, making the minimum investment $3B. These changes have hindered start-ups, making the semiconductor market less competitive. A good example of this has been how the existing chip makers have grown adept at controlling capacity by restricting investment. If this proves to be the case, the semiconductor market should be less cyclical in the future.

35.6 The Learning Curve

The learning curve has played a critical role in driving the rapid growth of semiconductor usage in electronics. It has a well-documented, long-term effect on semiconductor prices. The learning curve is simply the long-term erosion of manufacturing costs as a result of production experience, design shrinks, and yield increase. Cost decline as more devices are produced. This has the effect of making the supply curve downward sloping as shown in Figure 35.7. The effect is similar to a Veblin good, in which the demand curve slopes forward, driving prices up. In both cases, the market is fundamentally unstable. In the case of a downward sloping supply curve, the learning curve drives costs below price P1; suppliers are willing to provide devices for less than buyers are willing to pay. This drives the demand curve outward, thereby increasing production, lowering cost, and pushing the demand curve outward again. The effect has been an unceasing increase in demand as shown in Figure 35.8 and decrease in prices, as shown in Figure 35.9.

There are numerous technical driving forces that drive the learning curve. Learning economies occur as production increases and a plant becomes more efficient. As production experience is gained, a company literally learns how to make a product better: its designs are improved; die sizes are shrunk; and

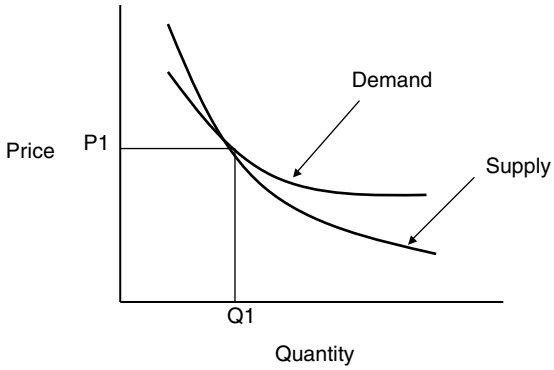


FIGURE 35.7 The downward sloping supply curve.

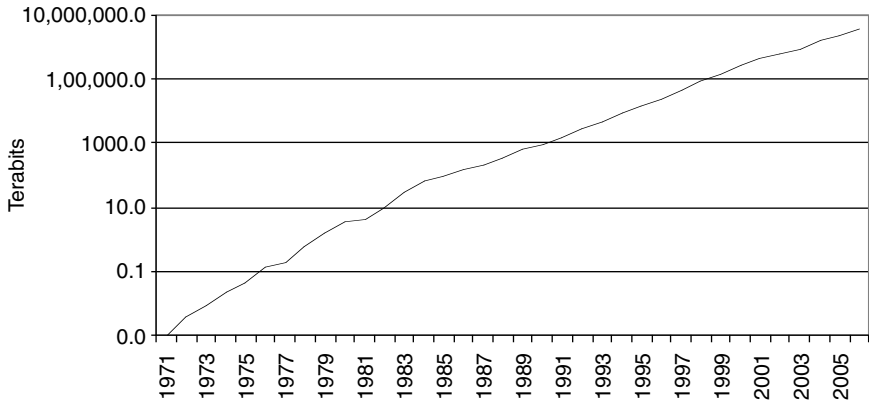


FIGURE 35.8 DRAM demand history.

its process is debugged. Consequently, wafer fab yields rise and even assembly yields rise. As yields rise, more product is shipped per unit of cost. Thus, revenues rise while total manufacturing costs stay constant or even declines, so average manufacturing cost per device declines.

Learning curve economies provide an important strategic advantage to the first company to enter a new market. The sooner a company enters a market, the quicker will be its gains from being first to ride down the learning curve. All things being equal, the first company to enter a market will have a manufacturing cost advantage throughout the market life of a product. This translates into a profitability advantage, which can be used to gain market share. The company must first be an efficient learner for this strategy to work. Otherwise, the comparative advantage provided by the learning curve will not offset the disadvantage of higher overall manufacturing costs. Korean companies proved this in competition against Japan in the early nineties, by exploiting new yield management techniques which, Japan's companies failed to recognize the benefits. (see Figure 35.10). Japan's early lead in yield gave its typical fab a \$750 M profit advantage over a Korean fab in 1992. But Korea reversed this in 1993, gaining a \$94 M profit advantage based on only a 6% yield advantage. It held this advantage through 1995, garnering another \$230 M.[5] The ultimate consequence of this was Japan's abandonment of the DRAM market, which it had fought so hard to dominate in the 1980s.

[5]Hutcheson, G. D. *Change in Chip Making and How it is Driving Process Diagnostics*. VLSI Research Inc., June, 1996.

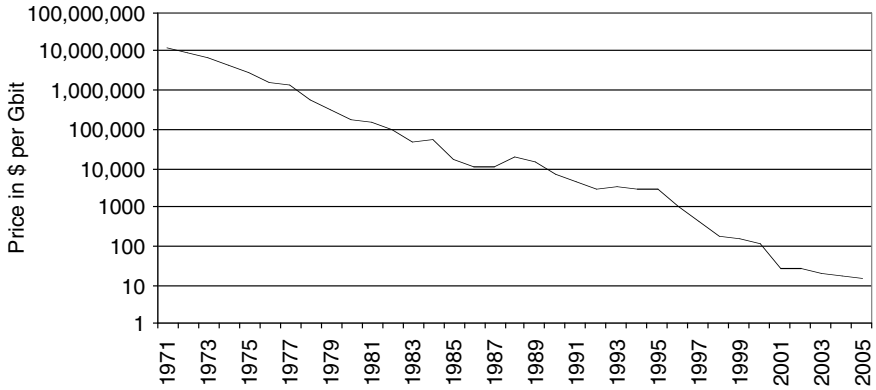


FIGURE 35.9 DRAM price history.

It was Thomas Hinkelman at Fairchild Semiconductor who first developed the strategy of using the learning curve to continually lower prices. This development in the mid-1960s allowed Fairchild to have a strong profit advantage while it eroded the profitability of its competitors. At the same time, lowering prices caused the early IC market to blossom.

There are three factors that are essential to a viable learning curve strategy. First, it is essential to have strength in component design and process technology. The first to design a new product is the first on the learning curve. Second, a top-notch sales and marketing ability is also essential. Sales and marketing expertise allows a company to quickly build a sales volume to get down the first part of the learning curve. A company must be able to quickly get customers to "design-in" a new device into their electronic equipment. Moreover, customer needs must be well understood in order to gain information about future device design requirements. Third, an expertise in manufacturing is needed. This allows a company to maintain its learning curve advantage over time. An example of how important these three strengths are can be seen in the competitive market structure that ensued after the first "Fairchild brain drain." The so-called brain drain occurred in the late 1960s when most of Fairchild's key talent left. Robert Noyce and Gordon Moore left to form Intel. With them, went most of Fairchild's design talent.

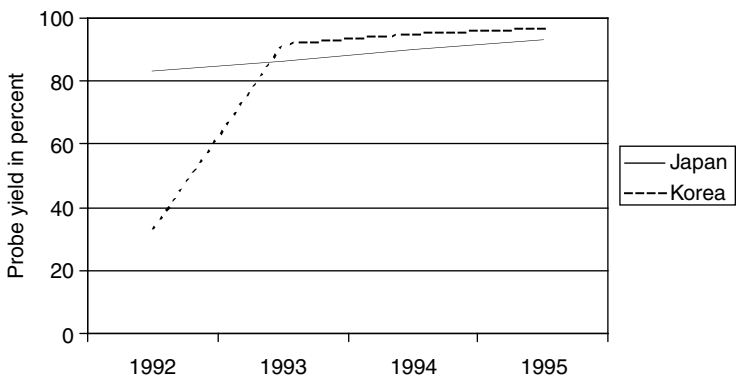


FIGURE 35.10 A learning curve case study: Korea vs. Japan in 4 Mb DRAMs.

Jerry Sanders left to form AMD. With Mr Saunders, went Fairchild's sales and marketing expertise. Charles Spork left to join National Semiconductor. With Mr Spork, went Fairchild's manufacturing strength.

A unique market structure emerged in the early 1970s as a result. Intel, having the design strength, was usually the first to enter a new market. Many of the most important semiconductor innovations of the 1970s emerged from Intel. Inventions such as the microprocessor and the DRAM-impacted everyday life were first made at Intel. AMD, would soon follow Intel into a market with its own version of the technology. AMD would often offer variations of the device that allowed it to win design-ins away from Intel. AMD eventually would catch-up and pass Intel on the learning curve. By this time, National Semiconductor would be entering the market. Its manufacturing prowess allowed it to immediately enter with a manufacturing cost advantage. Intel would soon exit the market for more fertile grounds with a new device generation. AMD would also eventually exit. This would leave only National. Intel, realizing its weaknesses, eventually became a leader in sales, marketing, and manufacturing. Correcting these deficits led to the industry giant that everyone knows today.

The competitive aspects of learning curve pricing cannot be ignored. There is a fragmentation of competition that occurs after a device has been introduced. The first firm into a market has a monopoly and prices accordingly. Once a few firms have entered the market, it will resemble an oligopoly. Prices remain high because the industry is still capacity limited at this point. A semiconductor manufacturer might charge $1000 for a part that costs only $50 to make.

Once a company can no longer protect its intellectual property (IP), the environment will resemble a purely competitive market. Once several companies have entered the market, no one has control over pricing. Prices fall to cost plus a reasonable profit. The market becomes a model of pure competition: the product is virtually homogenous with the exception that the producer's name is on it. However, this is more a legal formality. Pull the cover off a computer and one can often find the same part numbers from several different producers on the personal computer (PC) boards. Buyers seldom have brand preferences in this environment. There is also complete freedom of entry for existing firms into new product markets. This is especially true for commodity products such as DRAMs.

For those products with high development costs, a company can still "reverse engineer" another company's products. Process technology is usually available through equipment suppliers. There is also an abundant supply of venture capital to fund new start-up companies. Information about the market is easily obtained and advertising for existing product lines is rare. Advertising is usually limited to new product lines or image ads. Moreover, there has been little government intervention in the form of production or pricing controls as with agriculture. Consequently, learning curve pricing is merely the competitive response to a market that is in transition from a monopoly to pure competition.

35.7 The Technology Treadmill

Semiconductor's status as a commodity has been alternatively described as "the crude oil of industry in the 1980s" by Jerry Sanders, of AMD and as "the rice of industry" by the Japanese. Mr Sanders' statement was originally intended to imply that the semiconductor industry of the 1980s would reflect prosperity similar to that of the Organization of the petroleum exporting countries (OPEC) countries of the 1970s. However, while semiconductors are a commodity item much like crude oil, there was no cartel to support prices.[6] Rice may have been a more apt term because of the concept of a technology treadmill, which was first described by Cochrane in the 1950s. Cochrane's Theorem was originally developed to describe the impact of technology on agriculture. There are several parallels between semiconductors and agriculture that make this model valid. Many segments of the semiconductor market appear to be much like classic

[6]Ironically, OPEC's control over oil faded in the 1980s causing crude oil prices to fall like any commodity.

agricultural markets-for example, grade A red winter wheat. Like agriculture, mature semiconductor markets are almost textbook examples of pure competition.

One aspect of the semiconductor market, which is surprising similar to agriculture, is the manufacturing process. There are high capital costs in this market and these have grown at rapid rates just like in farming. Fixed costs are high, and most variable costs are non-recoverable once the process has been started. Materials, labor and capital must all be invested to make the wafer. There is no guarantee that this wafer will be good. The semiconductor manufacturer must "sow" its silicon, investing most of its money up front. Only later can a crop be yielded. Profitability is dependent on the absence of a "rain" of particles onto the wafer.

Moreover, the classic problems of oversupply in agriculture are also abundant in the semiconductor market. In agriculture, the problem of oversupply has been aggravated by a high degree of fixed assets, inelastic demand, low-income elasticity, rapid technological change, and competitive structure.[7]

To see the comparison, look at these parallels between semiconductors and farming as drawn from W.W. Cochrane's 1958 textbook, "Farm Prices, Myth and Reality." In this book, Cochrane explains a shift in the supply curve as being due to increasingly greater farm output occurring in combination with technological advancement. As new output-increasing technologies become available, they are adopted without changing any of the fundamental input variables such as plant and land. Moreover, the competitive structure of the agricultural industry accounts for a continued acceptance of output-increasing technology. He argues that this is typical of an inelastic market.

Cochrane's theory is clearly synonymous with what has taken place in the semiconductor industry. As more companies have entered the industry, competitiveness has grown keener. In order to remain competitive, all companies have been forced to adopt increasingly complex technologies. The next question then becomes, how does the market react to these increases?

Both agricultural and most semiconductor products are considered to be "generic" in the market place. Any one company's product can be substituted with that from any other. Therefore, product pricing becomes dependent upon market price. Market price, in turn, is dependent upon supply and demand. For example, if supply is less than demand, market prices will tend to be high. According to Cochrane, the only way for increasing income, in this type of market, is to adopt a cost-reducing technology. But the technology itself is output increasing. So, prices fall anyway. This causes a catch-22 situation for those not using the new technology. If they do not purchase the new technology, they will be squeezed out of the market. If they can afford the equipment, they will optimize output, thereby further forcing down the price of the commodity. In other words-the average farmer of the period was on a treadmill with respect to the technological advancements.

When prices dropped, farmers stayed in business even though they were losing money on total cost. Prices were below average total cost, but above average variable cost. The salvage value of assets was low. In this situation, the most profitable position is to stay in business and expand capacity. Next season, farmers would double up rows of planting and use more resources; each thinking that increasing production would increase last year's sales. Since supply increased rapidly but demand stayed constant, prices collapsed.

Similar conditions exist in the semiconductor market. Demand is inelastic in the short run. The result of this inelasticity is that prices drop rapidly once several competitors enter the market. The presence of several competitors creates a state of over-supply. Competitors respond by reducing die size. This increases production and lowers cost since better die can be made from the same area of silicon, which costs no more to produce. Moreover, fixed assets are more rapidly amortized. Supply increases, but demand stays constant and so prices drop. Income increases since production rises faster than the drop in prices. The cycle repeats itself over and over. The market is in continuous disequilibrium.

Notice that there is one important difference between semiconductors and agriculture. Semiconductors have a high degree of income elasticity, which aggravates supply by spurring an excess of entrants into the market during good times.

[7]Hathaway, D. E. "The Agricultural Treadmill," *Government and Agriculture*. 107–130. New York: Macmillan, 1963.

We all witnessed the effects as American farming went from small farms to giant land management corporations. Many of us in high technology are first-generation graduates from those farms. And this change in farming was propelled by technology-just like in the semiconductor industry. Small companies simply cannot afford to build huge DRAM plants. The current capital cost of a $1.5B plant is well beyond their reach. Venture capitalists cannot afford to fund this. Moreover, even major corporations find it difficult to stay in this market. They often quickly move to microprocessors and application-specific integrated circuits, where they can have some IP protection. Consequently, the commodity semiconductor market has evolved into the political preserves of a few nations.

35.8 Technological Driving Forces

In contrast to other industries, planning organizations in semiconductor manufacturing must be much more attuned to technological change. This is because a new semiconductor manufacturing plant can grow obsolescent in just 3 years time. It will usually be obsolete by its 15th birthday, without major upgrades. Companies must constantly renovate and rebuild manufacturing capacity. This technological pace plays a critical role in determining the success or failure of both semiconductor companies and equipment suppliers. The role that technology plays is felt in three key ways:

Technological driving forces start with a demand for new device technology and higher complexity. These two forces will drive changes in the physical, processing, and operational parameters of the device. Changes in these factors will, in turn, change the way in which devices are manufactured. New manufacturing requirements drive demand for new generations of equipment.

35.8.1 Technological Pace

This section shows how to derive equipment demand from the various technical forces that drive manufacturing. Conversely, the same approaches can be used to show how equipment either limits or expands the ability of the semiconductor industry to grow. For the purposes at hand, technological pace has been segmented into three basic types: technology's current state, evolutionary development, and revolutionary development. How each affects the market is examined below.

The current state of technology is the easiest to explain since it is the basis for an equipment market's size. There is almost a pure mathematical relationship that exists between the current state of technology and the size of any given equipment market. This relationship is founded on the basic premise that a company must produce a product before it can be sold. In turn, a company must have the capacity to produce the product.

The most basic measure of production in a wafer fabrication plant is the wafers itself. Packages, or die are the key measure of production in test and assembly. Total demand for equipment can be determined by tracking these measures of production.

Market demand for virtually any type of equipment can be determined by the equation shown below:

$$D = \frac{P \times S}{U} + R \tag{35.1}$$

where D, market demand in units; P, year to year change in production volume in units; S, number of steps per unit of production; U, annual system throughput in units; R, replacement.

The equation is a static state calculation of demand for 1 year. The way in which both evolutionary and revolutionary technical change affect the market, can also be derived from this equation by adding a time element.

Market growth is determined by two factors, one is market driven and the other is technically driven. Routine market growth in production drives an increase in P, the year-to-year change in production

volume. Technological change drives a multiplicative increase in *S*, the number of steps per unit of production. Thus, it can be seen that technological change can have a more powerful effect on market demand than routine market growth.

35.8.2 Evolutionary Technical Change

Evolutionary technical change has a major impact on accelerating market growth for semiconductor equipment. Evolutionary change can be characterized as those developments, which result from continued improvement of an existing technology. They tend to be time dependent and are reasonably predictable. For example, Moore's law is one example of evolutionary technical change. Steady improvements in line-width resolution, ion implantation, plasma etching, etc., have made these advances possible. The result has been a steady predictable growth in processing complexity.

The growth in process complexity has been well documented as being a key factor driving growth in equipment sales. As was mentioned previously, each additional step has a multiplicative effect on equipment demand. Evolutionary technical change is the primary driving force of growth in equipment sales.

35.8.3 Revolutionary Technical Change

Revolutionary technical change in manufacturing determines major shifts in demand from one type of equipment to another. It can be characterized by those developments, which result from technical breakthroughs. The nature of these breakthroughs is highly unpredictable and cannot be forecasted. They usually occur at times when the evolution of existing technology fails to progress. Typically, it runs into a key technical limitation that cannot be overcome. Consequently, a new and radically different technology emerges to take the place of the older technology.

The shift from projection aligners to wafer steppers in the early 1980s is a primary example of revolutionary technical changeover. Projection aligners went through four evolutionary generations in the mid-to-late-1970s. They were the dominant lithography tools in use at the time. Four to five micron resolution was routinely achievable by the mid-1970s. Sub-micron resolution was available in 1981 with the advent of deep ultra-violet (DUV) projection. However, overlay registration emerged as a new technical problem when the industry passed below two-micron geometries. The first generation of DUV projection aligners was not capable of achieving the required registration. The advent of the wafer stepper solved the registration problem, thereby allowing the industry to progress. This single technical limitation paved the way for the wafer stepper. The impact on market share was dramatic. The lithography market share held by wafer steppers rose from 7% in 1978 to 56% in 1984. Projection aligners' share of the market fell from 61% in 1978 to 28% in 1984. A similar change occurred in the nineties, with the changeover from steppers to scanners.

Revolutionary change usually has little measurable impact on overall market growth. It instead merely shifts capital from one technology to another. For example, in the total wafer exposure equipment market grew from $46.3 in 1975 to a 1980 peak year of $269.5 M—a compound annual growth rate of 42%. The compound annual growth between 1975 and the peak year of 1984 was 40%. The reason for this lack of additional market growth is that while steppers cost twice as much at one-half the throughput of a projection aligner, they achieved much higher yield. Consequently, the net cost per good die was actually lower. Thus, the change in overall market potential was not measurable. This is not to say that revolutionary technological change does not contribute to market growth. The semiconductor market would have stagnated without the advent of the wafer stepper. Revolutionary technological change allows, for the steady progression of market demand, by keeping the industry on Moore's curve. It also often causes major competitive shifts within markets when market leaders are asleep at the wheel. However, it seldom accelerates growth in equipment sales.

35.8.4 Characteristics of Technological Driving Forces

The ways in which technological change can occur is multifaceted. The various technology driving forces can be partitioned into four major paths for equipment development. These paths are: device technology, device complexity, processing parameters, and operational parameters. Each of these have several important sub-segments that are important drivers of equipment developments, which can be quickly multiplied into hundred's of sub-categories. These will again unfold into numerous rivulets of specialization. The possible combinations and permutations are virtually infinite. Consequently, it is necessary to limit the scope of any analysis to be able to comprehend the significant forces that drive technology. In the work that follows, technological driving forces will mainly be discussed according to how they pertain to processing parameters and operational parameters. These are generally the ones, which will more directly affect the economic effectiveness of manufacturing equipment. Device technology and complexity tend to serve as drivers of processing and operational parameters. Consequently, device technology and complexity will be largely ignored except as it relates to these parameters.

35.8.4.1 Processing Parameters

Processing parameters make most of their impact felt on wafer fabrication equipment. Changes in these parameters largely determine a need for new manufacturing equipment generations. They also provide a measure of the manufacturing prowess of a semiconductor company. Consequently, they are important business drivers for semiconductor manufacturers. There are five key parameters that are the most critical. They are:

- Wafer and die sizes
- Process steps
- Line-width
- Defect density
- Materials used.

Wafers are the one common denominator throughout the semiconductor industry. They provide a readily available yardstick against which both manufacturers and the industry can be evaluated. The revenues that a semiconductor manufacturer can generate from a square inch of silicon are an excellent measure of its marketing prowess. Its manufacturing prowess can be measured by its cost to produce that same square inch of silicon. These costs are largely controlled by wafer size and die size. Wafer sizes are continually increasing. So too are die sizes. Wafer size changes have been one of the most important driving forces of change in wafer fabrication equipment.

Equipment manufacturers will invariably place wafer size parameters high on their list of trends to watch. Seemingly minor changes in wafer size either diameter or thickness can wreak major design changes in new equipment. It immediately makes older equipment obsolete, thereby creating new demand. Consequently, from an equipment manufacturer's viewpoint, accurate prognosis of wafer size changes is of prime importance.

Die sizes are an important parameter because of their relevance to device complexity, yield, and manufacturing economies. The most simplistic of these relationships is how die size and wafer size interrelate. Each die is an IC. Consequently, the number die that can be put a wafer is a direct determinant of the sales that a wafer fab can achieve. This is why the term "Silicon Real Estate" is commonly used by industry veterans.

Silicon real estate is so critical for the simple reason that more die can be put on a wafer if they are made smaller, or conversely, if the wafer is made larger. Today's standard wafer size is 300 mm, or approximately 13 in. In the early days of the industry, wafers measuring 2 in. were the standard. A 13-in. wafer contains more than thirty times the silicon surface area of a 2-in. wafer. An 300 mm wafer that can produce as many as 100 of the same size die and at selling prices of $60.00 each can generate achieve revenues of $6000. In contrast, the 2-in. wafer could produce only two die and generate $120 of revenues.

However, equipment fitted out to handle 2-in. wafers would cost more than in a standard configuration for 300 mm. Even in standard configurations of 4-, 5-, 6-, or 8-in. wafers, there is no discount for opting for smaller wafers in equipment prices. Moreover, throughput—in wafers per hour—is often the same. Since, equipment depreciation is roughly half the cost of producing a wafer, chip manufacturers have almost always opted for the largest wafer size available when building a new facility.

The cost of manufacturing a wafer has typically increased 30% between each generation of size. This happens because the relative amount of resources needed to process a wafer is only loosely related to size. Moreover, once a new generation of equipment has been designed around a wafer size, the cost to equip a factory is roughly the same as that for equipment of the same CD generation built to handle smaller wafer sizes. So, a semiconductor manufacturer will almost always opt for a larger wafer size.

There is one exception to this rule: most prudent manufacturers will be hesitant about switching to the latest wafer size. The technology for the latest wafer size is seldom fully developed. Developing a new generation of wafer size can be very expensive for a semiconductor company. Equipment is much more expensive and less reliable, while yields are lower and material costs are higher for a new generation wafer size. Thus, net pretax profits are lower. Semiconductor manufacturers have typically bought equipment that is upgradable to the largest wafers possible; but actually process wafers that are one size smaller. Two advantages accrue as a result: first, it can readily expand to the larger wafer size when needed; second, yield will be higher, since equipment designed for larger wafers will have better uniformity on smaller wafers. The advantage for equipment industry is that manufacturers can focus their efforts on a common platform for a new CD generation. Developing two platforms adds unnecessary development costs, which will ultimately be burdened on the chip industry in the form of higher equipment prices.

Nevertheless, the real pressure on a semiconductor manufacturer is still to reduce die size, for this improves both output quantity and yield. Line-widths are a critical determinant of die size. Total output of die will improve as the square of the reduction in line-width. Moreover, yield will have an exponential effect on good die-outs. All that typically needed to accomplish a die shrink, within a major CD generation, is a new mask design with finer line-widths. Together these improvements accrue major advantages to a semiconductor manufacturer. So its emphasis will be upon die size reductions first, and upon wafer size increase later. These factors have been a key driving force for semiconductor manufacturers, since the beginning of the industry.

An example of how reduction in die size increases quantity can be seen in the evolution of the 4 K DRAM. When the 4 K DRAM was first introduced in 1971, it was produced on a die that was roughly 40,000 mile^2 in area. By 1979, this circuit had undergone four mask design changes and had been reduced to about 12,000 mile^2. But the product continued to be made on a 3-in. wafer. In return for that shrinkage, the same wafer size was producing 589 total die in 1979, vs. 176 total die in 1971. This represented a 3.3-fold increase in output due to just die shrinkage alone. Yield increase was equally dramatic. At the time, 40,000 mile^2 die would typically yield 20% net good die at wafer probe. A 12,000 mile^2 die would net about 60% good die. These combined advantages gave semiconductor manufacturers a 10-fold increase in output. And the only price paid was redesign and relayout of the die.

Later, the gains were even more staggering. The 16 Mb DRAM could hold just over 200 die per wafer when it was first introduced at 0.6-μm CDs. It was not long before the best manufacturers could put 1200 die on a wafer, using 0.2-μm CDs-a 6-fold increase. The yields also started (35%) and finished higher (95%) level-giving a 27-fold increase.

However, there is a downside to these gains: they make predicting capacity requirements immensely difficult, as an early generation 16 Mb factory would have produced only 15 M die per year, whereas a leading edge factory just a few years later could produce 257 M die. This is why static planning analyses by competitors have historically played such a key factor in causing gluts.

One might ask, is there an upper bound on practical limits to die size? The answer is a qualified yes. But it is largely limited to the state-of-the-art of technology, more than it is to any fundamental limitations. The reason has to do with defect density and yield. Since yield is limited largely by die area. Line-width and component count determine die area. Defect density is determined by the state of manufacturing technology.

This leads to the issue of packing density. Just how much can be packed in those chips? That will be dependent upon the size of the component (transistor, resistor or capacitor) put down on the die, as well as the line-width in-between each die. With current technology, there are three practical limits to packing density. One is the width and the current carrying capacity of the metal stripes connecting each element. Another is the voltage that can be withstood by each element. A third are the topographical layout limits that can be achieved with single layer, or with two layer, conducting surfaces. Transistors require three connections.

35.8.4.2 Operational Parameters

Operational parameters will impact manufacturers in four main areas. They are:

- Operating speed
- Pin count
- Package type
- Reliability.

There has been continuous emphasis on greater device speed over the past three decades. Device speed is an important business driver. For semiconductor manufacturers, higher speed means value added to their products. Not every die on a wafer is the same. Variances in manufacturing and defects can cause operating speed to vary by as much as two times. The fastest parts can bring from 100 to 200% price premium. Device speed is also an important business driver for both wafer fabrication equipment and test systems suppliers. Process uniformity and low contamination levels in process equipment yields higher speed devices. Consequently, the value added through higher device speed drives demand for cleaner process equipment and inspection equipment. Speed drives demand for greater timing accuracy in test equipment, as poor timing accuracy can result in high speed parts being binned at a lower grade or worse, being failed altogether.

35.9 Factory and Equipment Economics

The first economic analysis of the factories and equipment used to produce semiconductors was presented at ASEE in 1981.[8] It was a factory wide model that was developed to address the issue of declining profitability due to increasing capital investments. Three generations of factories were analyzed based on level of integration: small-scale integration/medium-scale integration, large-scale integration, and very large-scale integration (VLSI). The capital costs of each factory were approximately double that of the preceding generation, with VLSI factories costing a record $20 M for the equipment alone. The model was unique from a classic financial analysis in that it incorporated the technical capabilities, such as resolution and defect density, of each factory type to calculate cost per good device. It concluded that higher cost factories were actually more profitable for producing what were then state-of-the-art devices than lower cost, early generation factories, because they had higher yields, which resulted in lower cost per good device.

While the conclusions were proved out by a continued escalation in the cost of chip factories, the weakness of this model was that, with its hundred's of variables, it was too complex to be easily comprehended. Second, the industry was driven intuitively to the same conclusion by its own technical work and by market forces. Third, because factory models could be so complex they could be easily manipulated. This led to the development of simpler models that were focused on specific pieces of equipment. These are commonly referred to as cost of ownership or CoO models.

Today, almost all equipment selection and development are made with the aid of CoO model. CoO models are typically used to evaluate and compare tools prior to purchase. They can also be effective in

[8]Hutcheson, G. D. *VLSI—An Economist's Viewpoint*. ASEE, January 21, 1981. See also Hutcheson, G. D. *A Capital Investment Model for the Semiconductor Industry*, A Thesis Presented to The Faculty of the Department of Economics, San Jose State University, December, 1984.

the development of a tool to set design targets or to forecast the market acceptance of different technologies.

The industry's first CoO model was developed in the early 1980s to predict the future course of lithography tools.[9] At the time, there was a debate raging about which tool type was best: DUV projection, e-beam direct write, or steppers. Operational cost per hour was used to predict accurately that the stepper would emerge the victor in this hotly contested market. This model is similar to the earlier factory model that it incorporated the technical capabilities of each factory type to calculate cost per good device. It started by calculating all fixed and variable costs associated with producing wafers for each different type of tool using the following formula:

$$\mathrm{Co} = (E/S) + M \times T/N + \mathrm{Mt} + L + F \tag{35.2}$$

where Co, operational cost per hour; E, cost of equipment, including installation; S, serviceable life of the equipment in hours; M, materials cost per wafer per mask layer; T, net throughput per hour, including all utilization factors; N, number of mask layers; Mt, average cost of maintenance per hour of operation; L, direct and indirect labor cost per hour; F, floorspace cost per hour.

These costs were calculated for the serviceable life of the tool and then brought back to total cost per hour of operation. Once a device type was selected, the ideal die size for each type of tool was simulated using the device's component count and each tool's resolution and overlay capability. Then, mask levels, die size, and tool specific defect density were used to calculate yield using a Poisson distribution:

$$Y = \exp(-ADN) \tag{35.3}$$

where Y, yield; A, die area; N, number of mask layers.

Finally, net-good-die-per-hour was arrived at for each tool using net throughput (including utilization), die size, and yield:

$$\mathrm{Gd} = (T \times D \times Y \times L)/N \tag{35.4}$$

where Gd, good die per wafer; T, net throughput per hour, including all utilization factors; D, total die sites per wafer; Y, yield; L, line yield; N, number of mask layers.

This value was divided into total operational cost to arrive at cost per good die:

$$\mathrm{Cd} = \mathrm{Co}/\mathrm{Gd}$$

This model's strength was in its accuracy. It was adopted my some of the leading equipment manufacturers as an analytical tool to aid in selling equipment.[10] However, the fact that it needed a specific device design was a hindrance for both practical and political reasons. On a practical basis, most fabs run many designs at a time, making a design dependant model accurate on a per wafer basis or only on the largest fabs. On a political basis, the model relied heavily on DRAM designs as a vehicle for analysis because they were easy to model. But it wasn't until the late 1980s when the American semiconductor industry began pursue economic models as a way to benchmark itself against Japan's industry. By that time America's production of DRAMs was diminishing in consequence. So a design independent model that gave generic answers on a cost-per-wafer basis was needed.

In the mid-to-late 1980s, American semiconductor producers were came fire from more efficient Japanese producers. America was found to be weak in manufacturing and so SEMATECH was founded to

[9]Hutcheson G. D., and J. D. Hutcheson, *The VLSI Capital Equipment Outlook, Technical Ventures*. 2124–2132. October 1981 and Hutcheson, G. D. *Profitable Solutions to Lithography*, SEMICON/West Technical Program. 69–76. May 1982.

[10]GCA, TRE, and Perkin-Elmer used this model for lithography equipment in the early 1980s. It was then adapted for Perkin-Elmer to derive the cost-per-wafer of different etchers in 1983. In 1987, the model was again adapted by Novellus to market CVD equipment.

bring back manufacturing strength.[11] As this crisis came to a head, SEMATECH became the industry's champion for economic modeling. Early in the crisis, the industry focused on technology as a path to leadership. But, Paul Castrucci, then CoO of SEMATECH, believed that this needed to be balanced with an in-depth cost analysis of the tools in use and those in development. This effort ultimately led to the current semiconductor industry focus on CoO for production equipment.

Today, CoO models for production equipment have become an accepted way to aid in the purchasing decision for new production equipment. It differs from early models in that it calculates an equipment's equivalent cost per good wafer instead of its cost per good die. The most common CoO model in use today can be summarized by the following formula:

$$\text{CoO} = C/(T \times U \times S) \tag{35.5}$$

where C, equipment; T, throughput; U, utilization; S, serviceable life of the equipment.

Equipment costs include the sum of its initial purchase price and installation, plus the sum of the costs for consumable materials, maintenance, labor, and floorspace over the serviceable life of the tool. Throughput is defined as the maximum available per unit time. This includes only time associated with the loading, processing, and unloading of wafers. Utilization is the expected percentage of actual throughput over the life of the tool to its theoretical maximum throughput in the same period. The strength of this model is in its relative simplicity and focus on cost at the wafer level. This makes it relatively easy for people to agree in capital decision-making. Its weakness is the lack of inclusion of yield, which is often the most important factor driving competitiveness. Another weakness is that it focuses most of the attention on cost, as throughput and utilization are assumed to be relatively constant between chip manufacturers. This is seldom the case.

[11] SEMATECH's name was forged from the words semiconductor manufacturing technology in order to ensure its focus on manufacturing.

Appendix A: Physical Constants

Quantity	Symbol	Value	Units
Speed of light in vacuum	c	2.99792458×10^{8}	$\mathrm{m\,s^{-1}}$
Magnetic constant	μ_0	1.25663706×10^{-6}	$\mathrm{N\,A^{-2}}$
Electric constant	ε_0	8.85418782×10^{-12}	$\mathrm{F\,m^{-1}}$
Newtonian constant of gravitation	G	6.67310×10^{-11}	$\mathrm{m^3\,kg^{-1}\,s^{-2}}$
Planck constant	h	6.62606876×10^{-34}	J s
Elementary charge	e	1.60217646×10^{-19}	C
Electron mass	m_e	9.10938189×10^{-31}	kg
Proton mass	m_p	1.67262158×10^{-27}	kg
Neutron mass	m_n	1.67492716×10^{-27}	kg
Fine-structure constant	α	7.29735253×10^{-3}	
Avogadro's constant	N_A	6.02214199×10^{23}	$\mathrm{mol^{-1}}$
Molar gas constant	R	8.31447215	$\mathrm{J\,mol^{-1}\,K^{-1}}$
Boltzmann constant	K	1.38065032×10^{-23}	$\mathrm{J\,K^{-1}}$
Stefan–Boltzmann constant	σ	5.67040040×10^{-8}	$\mathrm{W\,m^{-2}\,K^{-4}}$
Wien's displacement law constant	B	2.89776865×10^{-3}	m K
Bohr magneton	μ_B	9.27400899×10^{-24}	$\mathrm{J\,T^{-1}}$
Compton wavelength	λ_C	2.42631022×10^{-12}	m
Electron magnetic moment	μ_e	-9.2847636×10^{-24}	$\mathrm{J\,T^{-1}}$

Appendix B: Units Conversion

Non-SI Unit	Factor	SI Unit
1 angstrom (Å)	1.0×10^{-10}	meter (m)
1 atmosphere (atm)	1.01325×10^{5}	pascal (Pa)
1 atomic mass unit (u)	1.66053873×10^{-27}	kilogram (kg)
1 British thermal unit$_{th}$ (Btu$_{th}$)	1.054350×10^{3}	joule (J)
1 calorie$_{th}$ (cal$_{th}$)	4.184	joule (J)
1 electronvolt (eV)	1.602177×10^{-19}	joule (J)
1 erg (erg)	1.0×10^{-7}	joule (J)
1 fluid ounce (floz)	2.957353×10^{-5}	cubic meter (m^3)
1 foot (ft)	0.3048	meter (m)
1 gallon (gal)	3.785412	liter (L)
1 gauss (Gs)	1.0×10^{-4}	tesla (T)
1 inch (in.)	2.54	centimeter (cm)
1 micron (μ)	1.0	micrometer (μm)
1 mile (mi)	1.609344	kilometer (km)
1 millimeter of mercury (mmHg)	133.3224	pascal (Pa)
1 ounce (oz) [avoirdupois]	28.34952	gram (g)
1 pound (lb) [avoirdupois]	0.4535924	kilogram (kg)
1 torr (Torr)	133.3224	pascal (Pa)
1 watt hour (W-h)	3600	joule (J)
1 year	3.155815×10^{7}	second (s)

International System of Units (SI) Prefixes

Prefix	Symbol	Factor
Yotta	Y	10^{24}
Zetta	Z	10^{21}
Exa	E	10^{18}
Peta	P	10^{15}
Tera	T	10^{12}
Giga	G	10^{9}
Mega	M	10^{6}
Kilo	k	10^{3}
Hecto	h	10^{2}
Deka	da	10^{1}
Deci	d	10^{-1}
Centi	c	10^{-2}
Milli	m	10^{-3}
Micro	μ	10^{-6}
Nano	n	10^{-9}
Pico	P	10^{-12}
Femto	f	10^{-15}
Atto	a	10^{-18}
Zepto	z	10^{-21}
Yocto	y	10^{-24}

Appendix C: Standards Commonly Used in Semiconductor Manufacturing

1. ANSI/ASQC Q91-1987 (ISO 9001) Quality Systems—Model for Quality Assurance in Design/ Development, Production, Installation, and Servicing.
2. ASTM D4327—Standard Test Method for Anions in Water by Chemically Suppressed Ion Chromatography.
3. ASTM E595—Testing and Measuring Outgas Rates.
4. European Union (EU) Directives—Three mandatory primary directives for semiconductor equipment:
   a. Machinery Directive (89/392/EEC, 93/44/EEC, 93/68/EEC),
   b. Electromagnetic Compatibility Directive (89/336/EEC, 92/31/EEC, 93/68/EEC), and
   c. Low Voltage Directive (73/23/EEC, 93/68/EEC).
5. Factory Mutual 7-7 Loss Prevention Data for Semiconductor Fabrication Facilities.
6. FED-STD-209E Airborne Particulate Cleanliness Classes in Cleanrooms and Clean Zones.
7. IES-RP-CC002-86 Laminar Flow Clean Air Devices.
8. IES-RP-CC001.3 HEPA and ULPA Filters.
9. IES-RP-CC006.2 Testing Cleanrooms.
10. IES-RP-CC007.1 Testing ULPA Filters.
11. IES-RP-CC021.1 Testing HEPA and ULPA Media.
12. IES-RP-CC024.1 Measuring and Reporting Vibrations in Microelectronic Factories.
13. ISO 14644-1 Cleanrooms and Associated Controlled Environments.
14. ISO 9000-1:1994 Quality Management and Quality Assurance Standards—Part 1: Guidelines for Selection and Use.
15. ISO 9000-2:1997 Quality Management and Quality Assurance Standards—Part 2: Generic Guidelines for the Application of ISO 9001, ISO 9002, and ISO 9003.
16. ISO 9000-3:1997 Quality Management and Quality Assurance Standards—Part 3: Guidelines for the Application of ISO 9001:1994 to the Development, Supply, Installation, and Maintenance of Computer Software.
17. ISO 9000-4:1993 Quality Management and Quality Assurance Standards—Part 4: Guide to Dependability Programme Management.
18. NFPA 318—Standard for the Protection of Clean Rooms.

19. NFPA 325M—Fire-Hazard Properties of Flammable Liquids, Gases, and Volatile Solids.
20. NFPA 49—Hazardous Chemicals Data.
21. NFPA 70—National Electrical Code (U.S.).
22. NFPA 704—Standard System for the Identification of the Fire Hazards of Materials.
23. NFPA 72—National Fire Alarm Code.
24. NFPA 79—Electrical Standard for Industrial Machinery.
25. OSHA 29 CFR 1910: Code of Federal Regulations, Title 29.
26. SEMATECH Application Guide 2.0 for SEMI S2 and S8.
27. SEMATECH Equipment Automation and Integration Test Users Group Recommendations.
28. SEMATECH Partnering for Total Quality—Standardized Supplier Quality Assessment (SSQA) Workbook, June 1994.
29. SEMATECH Software Quality—SEMATECH Supplier Standardized Quality Assessment (SSQA) Workbook, Module 3.
30. SEMATECH 92031014A-GEN Guidelines for Equipment Reliability, May 1992.
31. SEMATECH 93031567B-XFR Technician Training Guidelines: 1998.
32. SEMATECH 95082943A-ENG Software Process Improvement (SPI) Guidelines for Improving Software, Release 4.0, October 1995.
33. SEMATECH Technology Transfer Document 96103186A-TR Guidelines for Manufacturing Equipment Reference Manuals.
34. SEMATECH Technology Transfer Document #97063311E-ENG I300I Factory Guidelines, Version 4.1.
35. SEMATECH 97123411A-TR Software Quality Improvement Policy and Software Quality Key Indicators (4-Ups Metrics): Member Company Requirements for Suppliers.
36. SEMATECH Technology Transfer Document #98023468B-TR I300I Factory Guideline Compliance: Factory Integration Maturity Assessment for 300 mm Production Equipment, Version 4.0.
37. SEMATECH Technology Transfer Document #99033693A-ENG Integrated Minienvironment Design Best Practice.
38. SEMI E1.5 Standard for 150 mm Plastic and Metal Wafer Carriers, General Usage.
39. SEMI E1.7 Standard for 200 mm Plastic and Metal Wafer Carriers, General Usage.
40. SEMI E1 Specification for 3 in., 100, 125, and 150 mm Plastic and Metal Wafer Carriers.
41. SEMI E2 Specifications for Quartz and High-Temperature Wafer Carriers.
42. SEMI E2.2 (Reapproved 0299) Standard for 200 mm Quartz and High-Temperature Wafer Carriers.
43. SEMI E4 SEMI Equipment Communications Standard 1 Message Transfer (SECS-I).
44. SEMI E5 SEMI Equipment Communications Standard 2 Message Content (SECS-II).
45. SEMI E6 Facilities Interface Specifications Guideline and Format.
46. SEMI E7 (Reapproved 0699) Specification for Electrical Interfaces for the U.S. Only.
47. SEMI E10 Standard for Definition and Measurement of Equipment Reliability, Availability, and Maintainability (RAM).
48. SEMI E11 Guideline for 125, 150, and 200 mm Plastic and Metal Wafer Carrier Application.
49. SEMI E13 Equipment Communication Standard Message Service (SMS): Parts 1 and 2.
50. SEMI E14 Measurement of Particle Contamination Contributed to the Product from the Process or Support Tool.
51. SEMI E15 Specification for Tool Load Port.
52. SEMI E15.1 Provisional Specification for 300 mm Tool Load Port.
53. SEMI E26 Radial Cluster Tool Footprint Standard.
54. SEMI E30 Generic Model for Communication and Control of SEMI Equipment (GEM).
55. SEMI E31 Specification for Electrical Interfaces for Japan Only.
56. SEMI E33 Semiconductor Manufacturing Equipment Electromagnetic Compatibility.
57. SEMI E35 Cost of Ownership for Semiconductor Manufacturing Equipment Metrics.
58. SEMI E37 High-Speed SECS Message Services (HSMS) Generic Services.

59. SEMI E37.1 High-Speed SECS Message Services Single Session Mode (HSMS-SS).
60. SEMI E40 Standard for Processing Management.
61. SEMI E42 Recipe Management Standard: Concepts, Behavior, and Message Services.
62. SEMI E47.1 Provisional Mechanical Specification for Boxes and Pods Used to Transport and Store 300 mm Wafers.
63. SEMI E49.1 Guide for Tool Final Assembly, Packaging, and Delivery.
64. SEMI E49.2 Guide for High-Purity Deionized Water and Chemical Distribution Systems in Semiconductor Manufacturing Equipment.
65. SEMI E49.4 Guide for High-Purity Solvent Distribution Systems in Semiconductor Manufacturing Equipment.
66. SEMI E49.5 Guide for Ultrahigh Purity Solvent Distribution Systems in Semiconductor Manufacturing Equipment.
67. SEMI E49.8 Guide for High-Purity Gas Distribution Systems in Semiconductor Manufacturing Equipment.
68. SEMI E49.9 Guide for Ultrahigh Purity Gas Distribution Systems in Semiconductor Manufacturing Equipment.
69. SEMI E51 Guide for Typical Facilities, Services, and Termination Matrix.
70. SEMI E53 Event Reporting.
71. SEMI E54 Sensor/Actuator Network Standard Message Content.
72. SEMI E57 Provisional Mechanical Specification for Kinematic Couplings Used to Align and Support 300 mm Wafer Carriers.
73. SEMI E58 Automated Reliability, Availability, and Maintainability Standard (ARAMS): Concepts, Behavior, and Services.
74. SEMI E58.1 SECS-II Protocol for ARAMS.
75. SEMI E62 Provisional Specification for 300-mm Front-Opening Interface Mechanical Standard (FIMS).
76. SEMI E64 Provisional Specification for 300-mm Cart to SEMI El5.1 Docking Interface Port.
77. SEMI E72 Provisional Specification and Guide for 300 mm Equipment Footprint, Height, and Weight.
78. SEMI E78 Electrostatic Compatibility—Guide to Assess and Control Electrostatic Discharge (ESD) and Electrostatic Attraction (ESA) for Equipment.
79. SEMI E84 Specification for Enhanced Carrier Handoff Parallel I/O Interface Including Application Note.
80. SEMI E87 Specification for Carrier Management (CMS).
81. SEMI E89 Guide for Measurement System Capability Analysis.
82. SEMI E90 Standard for Substrate Tracking.
83. SEMI E94 Specification for Control Job Management.
84. SEMI E99 Carrier ID Reader/Writer Functional Standard: Specification of Concepts, Behavior, and Services.
85. SEMI E101 Provisional Guide for Equipment Front End Module (EFEM) Functional Structure Model (FSM).
86. SEMI F42 Test Method for Voltage Sag Susceptibility of Semiconductor Processing Equipment.
87. SEMI M1.8 Standard for 150 mm Polished Monocrystalline Silicon Wafers.
88. SEMI M1.9 Standard for 200 mm Polished Monocrystalline Silicon Wafers (Notched).
89. SEMI M1.15 Standard for 300 mm Polished Monocrystalline Silicon Wafers (Notched).
90. SEMI M8.12 Standard for 300 mm Polished Monocrystalline Silicon Test and Monitor Wafers (Notched).
91. SEMI S1 Safety Guideline for Visual Hazard Alerts.
92. SEMI S2 Safety Guidelines for Semiconductor Manufacturing Equipment.
93. SEMI S3 Safety Guidelines for Heated Baths.
94. SEMI S6 Safety Guideline for Ventilation.

95. SEMI S7 Safety Guideline for Environmental, Safety, and Health (ESH) Evaluation of Semiconductor Manufacturing Equipment.
96. SEMI S8 Safety Guidelines for Ergonomics Engineering of Semiconductor Manufacturing Equipment.
97. SEMI S9 Electrical Test Methods for Semiconductor Manufacturing Equipment.
98. SEMI S13 Safety Guidelines for Operation and Maintenance Manuals Used with Semiconductor Manufacturing Equipment.
99. SEMI T7 Specification for Data Matrix Wafer ID Marking.

Web Sites with Additional Information on Semiconductor Industry Standards

1. ANSI: www.ansi.org
2. ASTM: www.astm.org
3. Factory Mutual: www.fmglobal.com
4. IES: www.iest.org
5. ISO: www.iso.ch
6. NFPA: www.nfpa.org
7. OSHA: www.osha.gov
8. International SEMATECH: www.sematech.org

Appendix D: Acronyms

Acronym	Definitions
AAPSM	Alternating aperture phase-shift mask
AAS	Atomic absorption spectroscopy
ACLV	Across-chip linewidth variation
ACS	American Chemical Society
ADC	Automated defect classification
ADC	Automatic diameter control
ADSL	Asymmetric digital subscriber line
AEM	Analytical electron microscopy
AES	Auger electron spectroscopy
AFM	Atomic force microscopy
AGV	Automated guided vehicle
ALF	Advanced lithography facility
AMHS	Automatic material handling system
ANSI	American National Standards Institute
APC	Advanced process control
APCVD	Atmospheric pressure chemical vapor deposition
APS	American Physical Society
ARC	Anti-reflection coating
ARDE	Aspect ratio-dependent etching
ARL	Average run length
AS/RS	Automated storage and retrieval systems
ASET	Association for Super-advanced Electronics Technologies
ASIC	Application-specific integrated circuit
ASTM	American Society for Testing and Materials
ATE	Automated test equipment
ATPG	Automatic test pattern generation
AVS	American Vacuum Society
AWLV	Across-wafer linewide variation
BED	Boron-enhanced diffusion
BEOL	Back end of line
BGA	Ball grid array
BHF	Buffered HF
BiCMOS	Bipolar CMOS
BIST	Built-in self test
BO	Born–Oppenheimer
BPSG	Borophosphate silicate glass
BST	Barium strontium titanate

(continued)

Acronym	Definitions
BTS	Bias-temperature stress
CAD	Computer-aided design
CARRI	Computerized assessment of relative risk impacts
CBE	Chemical beam epitaxy
CCD	Charge-coupled device
CD	Critical dimension
CDA	Clean dry air
CD-AFM	Critical dimension atomic force microscope
CDMA	Code divided multiple access
CFM	Contamination-free manufacturing
CIM	Computer-integrated manufacturing
CIVA	Charged induced voltage alteration
CMOS	Complementary metal-oxide-silicon
CMP	Chemical–mechanical polishing
CoC	Cost of consumables
COG	Chrome on glass
CONWIP	Constant WIP
CoO	Cost of ownership
COP	Crystal-originated pits
CP	Cell projection
CPAA	Charged particle activation analysis
CRRC	Contact retaining ring carrier
CTD	Carpal tunnel dysfunction
CTE	Coefficient of thermal expansion
CV	Capacitance-voltage
C-V	Capacitance-voltage
CVD	Chemical vapor deposition
CZ	Czochralski
DAC	Digital-to-analog converter
DARC	Dielectric anti-reflective coating
DBCBT	Driven blade cantilever beam testing
DCS	Dichlorosilane
DD	Dual damascene
DESIRE	Diffusion-enhanced silylated resist
DF	Dielectric functions
DFT	Design for test
DFT	Density functional theory
DFT	Design for testability
DHF	Dilute HF
DIBL	Drain-induced barrier lowering
DIP	Dual in-line package
DL	Defect level
DLC	Diamond-like carbon
DLTS	Deep-level transient spectroscopy
DMA	Dynamic mechanical analysis
DMAH	Dimethyl aluminum hydride
DMSDMA	Dimethylsilyl dimethylamine
DNP	Distance to the neutral point
DNQ	Diazonaphtoquinone
dpm	Defective parts per million
DRAM	Dynamic random access memory
DRS	Diffuse reflection spectroscopy
DSP	Digital signal processor
DSW	Direct step on wafer
DUT	Device under test
DUV	Deep ultraviolet

(continued)

Acronym	Definitions
DZ	Denuded zone
EAPSM	Embedded attenuating phase-shift mask
EBIC	Electron-beam-induced current
EBL	Electron-beam lithography
ECR	Electron cyclotron resonance
ECS	Electro Chemical Society
EDS	Energy dispersive x-ray
EDTA	Ethylenediaminetetraacetate
EDX	Energy dispersive x-ray
EDXS	Energy dispersive x-ray spectroscopy
EELS	Electron energy loss spectroscopy
EEPROM	Electrically erasable programmable read only memory
EFA	Evolving factor analysis
EFR	Early failure rate
ELT	Edge lift-off testing
EM	Electromigration
EMI	Electromagnetic interference
EOR	End-of-range
EPROM	Erasable programmable read only memory
ERD	Elastic recoil detection
ESCA	Electron spectroscopy for chemical analysis
ESD	Electro static discharge
ESS	Environmental stress screening
EUV	Extreme ultraviolet
EUVL	Extreme ultraviolet lithography
EWFA	Evolving window factor analysis
EWMA	Exponentially weighted moving average
FA	Failure analysis
FCFS	First come first served
FD SOI	Fully depleted SOI
FDC	Fault detection and classification
FDDI–LAN	Fiber distributed data interface–local area network
FDMA	Frequency division multiple access
FEA	Finite element analysis
FEOL	Front end of line
FESEM	Field emission scanning electron microscope
FET	Field effect transistor
FFT	Fast Fourier transform
FIB	Focused ion beam
FIFO	First in first out
FMCS	Facility management control system
FMEA	Failure mode and effects analysis
FMI	Fluorescent thermomicrographic imaging
FOUP	Front opening unified pod
FPD	Flow pattern defects
FRACAS	Failure reporting and corrective action system
FRB	Failure review board
FSB	Fixed shaped beam
FSG	Fluorine-doped silicate glass
FSMCT	Fluctuation smoothing policy for mean cycle time
FSVL	Fluctuation smoothing policy for variance of lateness
FT-IR	Fourier transform infrared
FTPL	Fourier transform photoluminescence spectroscopy
FWI	Full wafer imaging
GAE	Gas-assisted etching
GC	Gas chromatography

(continued)

Acronym	Definitions
GILD	Gas immersion laser diffusion
GO	Gate oxide
GOI	Gate-oxide integrity
GWP	Global warming potential
HAST	Highly accelerated stress test
HCI	Hot-carrier injection
HDP	High-density plasma
HDP CVD	HDP chemical vapor deposition
HF	Hydrofluoric acid
HFCs	Hydrogen-containing fluorocarbons
HFOCs	Hydrogen and oxygen containing fluorocarbons
HIBS	Heavy ion backscattering
HMDS	Hexamethyldisilizane
HREM	High-resolution electron microscopy
HRTEM	High-resolution transmission electron microscopy
HSQ	Hydrogensilsesquioxanes
HVAC	High vacuum
IC	Integrated circuit
IC	Ion chromatography
ICDD	International Centre for Diffraction Data
ICP	Inductively coupled plasma
ICP-MS	Inductively coupled plasma mass spectrometry
ID	Inner diameter
IEST	Institute for Environmental Sciences and Technology
IFR	Intrinsic failure rate
ILD	Inter-layer dielectric
ILDs	Inter-level dielectrics
IMD	Inter-metal dielectrics
IMEC	Interuniversity Microelectronics Center
I-MR	Individuals-moving range
IP	Image placement; intellectual property
IPL	Ion projection lithography
I-PVD	Ionized physical vapor deposition
IR	Infrared
ISO	International Organization for Standardization
ISPM	In situ process monitor
ISS	Impulsive stimulated scattering
ISTS	Impulsive stimulated thermal scattering
ITRS	International technology roadmap for semiconductors
LAN	Local area network
LBFS	Last buffer first served
LCC	Life cycle cost
LDD	Lightly doped drain
LECIVA	Low-energy charge-induced voltage alteration
LES	Line end shortening
LIVA	Light-induced voltage alteration
LOCOS	Local oxidation of silicon
LPCVD	Low-pressure chemical vapor deposition
LPDs	Light point defects
LR	Learning rate
LSI	Large-scale integration
LST	Laser-scattering tomography
LTO	Low-temperature oxide
LWNQ	Least work next queue
MBE	Molecular beam epitaxy
MBPC	Model-based process control

(continued)

Acronym	Definitions
MCM	Multi-chip module
MCNC	Microelectronic Center of North Carolina
MCS	Material control system
MCT	Mercury cadmium telluride
MCZ	Magnetic field-applied Czochralski growth
MD	Molecular dynamics
MEEF	Mask error enhancement factor
MEF	Mask error factor
MEMS	Micro electrical mechanical systems
MES	Manufacturing enterprise system
MES	Manufacturing execution system
MFC	Mass flow controllers
MIAE	Mean integral absolute error
MIMO	Multiple-input, multiple-output
MISE	Mean integral squared error
ML	Multi-layer
MLC	Multi-layer ceramic
MLI	Multi-level interconnect
MOCVD	Metal organic chemical vapor deposition
MOS	Metal-oxide semiconductor (transistor)
MOSFET	MOS field effect transistor
MPC	Mask protective covers
MRS	Materials research society
MSE	Mean-squared error
MTBF	Mean time between failures
MTTR	Mean time to repair
MW	Microwave
Mw	Molecular weight
NAA	Neutron activation analysis
NFPA	National Fire Protection Association
NGL	Next generation lithography
NHE	Normal hydrogen electrode
NIST	National Institute of Standards
NMOS	N-channel metal-oxide semiconductor (transistor)
NRA	Nuclear reaction analysis
NRRC	Non-contact retaining ring carrier
NSOM	Near field scanning optical microscope
NTD	Neutron transmutation doping
NTRS	National Technology Roadmap for Semiconductors
OBIC	Optical beam-induced current
OBIRCH	Optical beam-induced resistance change
OED	Oxidation-enhanced diffusion
OEE	Overall equipment effectiveness
OEM	Original equipment manufacturer
OES	Optical emission spectroscopy
OHV	Overhead hoist vehicle
OLIC	Open-loop lamp intensity control
OPC	Optical proximity correction
OPP	Optical precipitate profiler
ORD	Oxidation-retarded diffusion
ORTC	Overall roadmap technology characteristics
OSQ	Organosilsesquioxanes
OSHA	Occupational Safety and Health Administration
PACE	Plasma-assisted chemical etching
PAG	Photo-acid generator
PCC	Predictor–corrector control

(continued)

Acronym	Definitions
PCD	Photoconductance decay
PCW	Process cooling water
PD SOI	Partially depleted SOI
PDF	Probability density function
PDS	Particle detection system
PEB	Post-exposure bake
PECVD	Plasma-enhanced chemical vapor deposition
PEEC	Partial element equivalent circuit
PFC	Perfluoro compound
PGILD	Projection GILD
PGV	Personnel guided vehicles
PIC	Physical interfaces and carriers
PICA	Picosecond imaging circuit analysis
PID	Proportional integral derivative
PL	Photoluminescence
PLASMEX	Plasma mechanical activation extraction of particulate contamination
PM	Preventive maintenance
PMD	Pre-metal dielectric
PMMA	Polymethyl methacrylate
PMOS	P-channel metal-oxide semiconductor (transistor)
PMT	Photo multiplier tube
PNP	P-doped N-doped P-doped (bipolar transistor)
PO	Purchase order
PO	Poly (propyleneoxide)
POA	Post-oxidation anneal
POTWs	Publicly owned treatment works
POU	Point-of-use
PPM	Parts per million
PPQ	Poly (phenylquinoxaline)
PRAT	Product reliability acceptance test
PREVAIL	Projection exposure with variable axis immersion lenses
PRIME	Positive resist image by dry etching
PSA	Product sensitivity analysis
PSE	Product-specific emulation
PSG	Phosphorus-doped silicon glass
PSM	Phase-shift mask
PTFE	Polytetrafluoroethylene
PULSE	Picosecond ultrasonic laser sonar technology
PVD	Physical vapor deposition
PWB	Printed wiring board
PXLA	Proximity x-ray Lithography Association
QBD	Q(charge)-to-breakdown
QFP	Quad flat package
QTAT	Quick turn-around time
RAM	Random access memory
RBS	Rutherford backscattering spectrometry
RET	Resolution enhancement technique
RF	Radio frequency
RFQ	Request for quote
RGA	Residual gas analysis
RGV	Rail-guided vehicle
RH	Relative humidity
RIE	Reactive ion etch
RLC	Resistance, L (inductance), and capacitance
RMS	Root mean square
RTA	Rapid thermal annealing

(continued)

Acronym	Definitions
RTCVD	Rapid thermal chemical vapor deposition
RTN	Rapid thermal nitridation
RTO	Rapid thermal oxidation
RTP	Rapid thermal processing
RTS	Rapid thermal silicidation
S/D	Source/Drain
SABRE	Silicon-added bilayer resist
SAC	Self-aligned contact
SAHR	Silylated acid hardened resist
SAM	Scanning acoustic microscopy
SCALPEL	Scattering with angular limitation projection electron-beam lithography
SE	Spectral ellipsometry
SECS	Semiconductor equipment communications standard
SEI	Seebeck effect imaging
SEM	Scanning electron microscopy
SEMI	Semiconductor Equipment and Materials International
SFQD	Site flatness deviation
SFQR	Site flatness range
SIA	Semiconductor Industry Association
Si-CARL	Silicon chemical amplification of resist lines
SILC	Stress-induced leakage current
SIMS	Secondary ion mass spectroscopy
SLC	Surface layer circuitry
SM	Scheduled maintenance
SMIF	Standard mechanical interface
SMT	Surface mount technology
SNR	Signal-to-noise ratio
SOD	Spin-on dielectric
SOG	Spin-on glass
SOI	Silicon on insulator
SOI CMOS	See SOI and CMOS
SOM	Sulfuric acid ozone mixture
SONET	Synchronous optical network
SPC	Statistical process control
SPIE	Society for Photo-optical Instrumentation Engineering
SPM	Scanning probe microscopy
SPT	Shortest process time
SPV	Surface photovoltage
SRAM	Static random access memories
SRC	Semiconductor research corporation
SSO	Simultaneous switching output
STI	Shallow trench isolation
STM	Scanning tunneling microscopy
TCAD	Technology computer-aided design
TDDB	Time-dependent dielectric breakdown
TDR	Time domain reflectometry
TED	Transient-enhanced diffusion
TEM	Transmission electron microscopy
TEM	Transverse electromagnetic (wave)
TF	Time-to-failure
TFAA	Trifluoroacetic anhydride
TIR	Total indicated reading
TIVA	Thermally induced voltage alteration
TMAH	Tetramethylammonium hydroxide
TMDS	Tetramethyl disilizane
TOF-SIMS	Time-of-fight secondary ion mass spectrometry

(continued)

Acronym	Definitions
TPM	Total productive maintenance
TQM	Total quality management
TRI	Toxic release inventory
TTL	Through the lens
TTM	Through the mask
TTT	Time–temperature transformation
TTV	Total thickness variation
TVS	Triangular voltage sweep
TXRF	Total reflection x-ray flourescence
UAMOM	UA method of moments
UBM	Under-bump metallization
UHV	Ultra high vacuum
ULSI	Ultra large scale integration
UPW	Ultra pure water
USG	Undoped silicate glass
UV	Ultraviolet
VAE	Variable angle ellipsometry
VASE	Variable angle spectral ellipsometry
VBD	Voltage-to-breakdown
VIS	Visible light
VLSI	Very large scale integration
VOC	Volatile organic compound
VPD	Vapor phase decomposition
VPD-DC	Vapor phase decomposition-droplet collection
VSB	Variable-shaped beam
W-CVD	Tungsten (W) chemical vapor deposition
WD	Wavelength dispersive
WDS	Wavelength dispersive spectroscopy
WECO	Western Electric Company
WIDNU	Within die non-uniformity
WIP	Work in progress
WIP	Work in process
WIWNU	Within wafer non-uniformity
WSW	Wafer starts per week
XPS	X-ray photoelectron spectroscopy
XRC	X-ray rocking curve
XRD	X-ray diffraction
XRF	X-ray fluorescence
XRIS	X-ray image sensor
XRL	X-ray lithography
XRT	X-ray tomography

Index

Ψ-MOSFET, **4**-26–27

A

Aberrations, optical lithography and, **18**-7–10
Abnormality detection, process control and, **23**-18–20
Abnormality process control, **23**-4
Abrasives, **17**-17–18
Absorbtivities, **11**-15–27
Accelerated testing, **30**-2–3
Acceleration factor, **30**-10–11
Acceleration voltage, **20**-25
Acceleration
  DC type, **7**-50
  ion implantation equipment and, **7**-49–58
  radio frequency linear, **7**-50–51
  tandem, **7**-51–52
Acoustic composition measurement, **25**-22–26
  sensor
    calibration, **25**-24–25
    configurations, **25**-23–24
    integration, **25**-25
  theory, **25**-22–23
Acoustic measurement method, interconnect film thickness and, **24**-39–44
  resolution and precision, **24**-44–45
Acoustic wafer temperature sensor, **25**-9–10
Acronyms, **D**-1–8
Additive behavior, filling and, **16**-28
Additive disturbance assumption, **23**-29–30
Advanced process control, **23**-5–7
Advanced wafer engineering, **4**-18–24
  crystal orientations, **4**-18–20
  Ge on insulator, **4**-23–24
  on-insulator substrates, **4**-23–24
  strained silicon-on-insulator, **4**-20–23
Aerial image analysis, **20**-66–67
AFM CD metrology, **20**-52–53
AFM nanomatching mask repair, **20**-62–63
Air, low-κ, **21**-56–57
ALD. *See* Atomic layer deposition.
ALD high κ materials, **13**-52–55
  metal oxide, **13**-52–55
ALD metal, **13**-80–82
ALD process characterization, **13**-9–13
  exposure time, **13**-10
  purge time optimization, **13**-10
  temperature, **13**-10–13
ALD reaction mechanisms, **13**-7–8
ALD reactor design, **13**-21–22
ALD transition, **13**-76–77
ALD W nucleation process, **13**-63–67
Alignment, optical lithography and, **18**-10–12
Alpha particles, **31**-6–8
Alt. PSM. *See* alternating phase shift mask.
Alternating phase shift mask (Alt. PSM), **20**-7–13
  fabrication of, **20**-8–10
  halftone phase shift mask, **20**-13–14
  layout consideration, **20**-10
  special cases, **20**-10–13
Aluminum, **13**-78–80
Aluminum CMP, **17**-46
Ammonia based silicon oxynitride films, **9**-12–13
Amorphization, **7**-28–30
Analog design, photomask device specific reticle issues and, **20**-71
Anneal behavior, **16**-34–36
Annealed hydrogen-implanted silicon wafers, **4**-11–12
Annealing, **11**-92
  copper, **11**-97
  ion implantation damage, **11**-85–92
  millisecond, **11**-88–90
  pulsed laser, **11**-90–91
  solid phase epitaxy, **11**-91
  spike, **11**-87–88
  Ta_2O_5 integration and, **13**-44–45
Annealing ambient, **9**-5
Anode film, **16**-18–19
Anomalous thermal expansion
  nickel silicides and, **10**-31–32
  unit cell dimensions, **10**-31
Aqueous solutions, **5**-4–7
  complexing, **5**-5–6
  oxidation state, **5**-7
  particle adhesion, **5**-8–9
  particle removal, **5**-8

pH, **5**-6–7
solubility, **5**-5–6
ARC films, **15**-10
ArF materials, **19**-40–46
post, **19**-51–53
transparent polymer systems, **19**-41–45
ArF transparent polymer systems
deprotection kinetics, **19**-43
exposure step, **19**-42–43
extending, **19**-45–46
line edge roughness, **19**-43
new classes of, **19**-44
new immersion fluids, **19**-49–50
typical type, **19**-41
ASIC design, photomask device specific reticle issues and, **20**-70
Aspect ratio, **21**-48–50
dependent etching, **21**-15
etch stop, **21**-35–36
Atomic layer deposition (ALD), **13**-4, **14**-1–33
applications, **14**-25–32
dielectrics, **14**-28–32
higher-*k* oxide capacitors, **14**-25–32
interconnect, **14**-31–32
metal gates, **14**-28–32
capacitor, **14**-2
chemical processes, **14**-5–20
illustrating of, **14**-6–8
sequential self-limiting processes, **14**-5–6
types, **14**-8–12
gate, **14**-2
interconnect, **14**-3
origins of, **14**-3–5
reactions, types, **14**-8–12
thermal, **14**-8–10
system technology, **14**-20–24
batch systems, **14**-24
single wafer, **14**-23–24
Auger, **29**-19–20
Automatic diameter control-induced perturbation, **3**-12
Availability metrics, **22**-9
Axial dopant distribution, **3**-7–8

B

Back end line etch process, **21**-55–57
Back end of line etch, **21**-40–46
Backside probing, **29**-13–14
Bake exposure step, **19**-31–37
Band edge measurement techniques, **25**-7
Band edge thermometry limitations, **25**-7–8
Barium strontium titanate, **13**-47–52
integration issues, **13**-52
microstructural effects, **13**-51–52
precursors, **13**-47–51
Barriers, tungsten CVD film and, **13**-62
Batch systems, ALD system technology and, **14**-24
Beam residual gas interactions control, **7**-61–65
Beam scanning system, **7**-52–58
dual, **7**-55–57
electromagnetic type, **7**-52–55
electrostatic type, **7**-52–55
Beams
acceleration voltage, **20**-25
imaging impact of, **20**-22–23
space charge effect, **20**-25
types, **20**-22–23
point, **20**-22–23
shaped, **20**-22–23
Bernas sources, **7**-42–45
BESOI. *See* bond and etchback SOI.
Bethe–Bloche regime, **7**-3
Black diamond low κ dielectric film, **13**-39–40
Bond and etchback SOI, **4**-4, **4**-13–14
Boron doped silicon glass, **13**-30–31
Boron, **31**-9–10
Borophosphosilicate glass, **13**-28–30
BOX, Si thickness measurements and, **4**-24–26
Bridging, **10**-27–28
Brightfield inspection, **27**-21–22
Buffered hydrofluoric acid, **5**-15–16
Build reliability, **22**-18–19
Bulk organic removal, **5**-18
Bulk trap, **9**-21–22
Buried layers, **7**-32–33

C

Calibration
critical dimension measurement and, **24**-4–17
IP metrology and, **20**-55
Capacitance voltage profiling, **28**-7–9
Capacitance, **32**-10–11
Capacitively coupled etching, **21**-10–11
Capacitor, **14**-2
Capacity, semiconductor manufacturing and, **35**-6–8
CAR negative tone, **20**-42
Carbon based low κ dielectric film, **13**-40
Carrier concentration, electrical characterization techniques, **28**-2–12
Carrier design, **17**-9–14
Carrier generation based semi global techniques, **29**-10
Carrier lifetimes
electrical characterization techniques and, **28**-19–23
generation, **28**-22–23
recombination, **28**-21–22
Carrier mobility, stress enhanced, **24**-35–36
Carrier, wafer fab logistics and, **33**-3–10
Catalytic ALD, **14**-11–12
CD metrology, **20**-48–53
AFM, **20**-52–53
optical, **20**-48–49
SEM, **20**-49–52
CD SEM (critical dimension scanning electron microscope), **24**-5–10
calibration, **24**-10–12
focus on, **24**-10–12
manufacturability, **24**-10–12
tool matching, **24**-10–12
Chamber walls, **21**-27–30

Channel crack failures, **12**-3–8
extracting materials properties, **12**-4–8
film thickness effects, **12**-3–4
Channel cracking, **12**-16–20
diffusion studies, **12**-18–20
Channel doping issues, **1**-28–36
Channel hot carrier injection, **30**-28–30
Channel length, **28**-14–15
Channel thickness, **4**-34–36
Channeling, **4**-33–34, **7**-5–6, **7**-20–24
Characterizations
high spatial resolution imaging, **28**-30–39
physical and chemical, **28**-30–62
auger electron spectroscopy, **28**-53–54
dopants, **28**-39–46
impurities, **28**-39–46
physical defects, **28**-54–62
Rutherford backscattering spectrometry, **28**-46–49
stress defects, **28**-54–62
total reflection X-ray fluorescence, **28**-49–51
X-ray fluorescence, **28**-49–51
X-ray photoelectron spectroscopy, **28**-51–53
techniques, **28**-1–62
electrical, **28**-2–30
Charge coupled device array spectrometers, **25**-12–14
Charged based measurements, **28**-23–25
Charging based global techniques, **29**-9
Charging, **21**-17
Charging, ion implantation process control and, **7**-65–68
Chemical characterization, **28**-30–62
auger, **29**-19–20
failure analysis and, **29**-18–21
Fourier transform infrared spectroscopy, **29**-20
secondary ion mass spectroscopy, **29**-20
X-ray analysis, **29**-19
Chemical composition, low-*k* dielectrics and, **12**-7
Chemical effects
chemical mechanical polishing and, **17**-36–42
copper CMP and, **17**-40–42
oxide dielectric CMP and, **17**-36–37
tungsten CMP and, **17**-37–39
Chemical etching, **21**-8–9, **3**-46–47
Chemical mechanical polishing (CMP), **17**-1–48
applications of, **17**-42–46
dielectric, **17**-42–44
ILD planarization, **17**-43–44
metal, **17**-44–46
polysilicon, **17**-44
chemical effects in, **17**-36–42
cleaning, post, **17**-12–13
damascene processing, **17**-3–4
edge effect, **17**-28–29
equipment, **17**-6–24
carrier design, **17**-9–14
classification of, **17**-7–14
endpoint methods, **17**-13–14
equipment, table motion, **17**-7–9
pad condition design, **17**-11–12
pads, **17**-15–16
polish tool integration, **17**-12–13
post cleaning, **17**-12–13
film removal, **25**-40–41
grinding, **17**-2
history of, **17**-4–6
lapping, **17**-2
mechanics of, **17**-25–28
mechanisms, **17**-25–42
non-patterned wafers, **17**-25–29
pads, **17**-15–16
patterned wafers, **17**-29–36
planarization, **17**-2
platen, **17**-2
polishing, **17**-2
post cleaning, **17**-46
dielectrics, **17**-46–48
metals, **17**-46–48
process control, **24**-47–49
film flatness and quality, **24**-47–49
random defects and, **27**-5–6
slurries, **17**-16–24
Chemical undercutting, **5**-9–10
Chemical vapor deposition, **13**-1–82
Chemically amplified relief image formation, **19**-23–40
exposure step, **19**-25–41
overview, **19**-23–25
Chemically amplified resists (CAR), **20**-42
negative tone, **20**-42
positive tone, **20**-42
Chemistry
CVD and, **13**-2–5
plasma and, **21**-4–10
Chip fabrication, **20**-30–31
Chrome etching, **20**-43–46
cleaning processes, **20**-45–46
dry, **20**-43–44
resist stripping and cleaning, **20**-45–46
wet, **20**-43
Chrome, **20**-40
Cleaning
CMP and, **17**-46
CVD chemistry and, **13**-4–5
processes
chrome etching and, **20**-45–46
supercritical carbon dioxide and, **6**-6–18
wafer preparation and, **3**-48–49
Cluster analysis, **26**-2–4
CMOS (complementary MOS devices), **1**-1
CMOS elevated source/drain, **3**-55–56
CMOS etch process modules, **21**-18–46
back end of line etch, **21**-40–46
damage considerations, **21**-39–40
dual inlaid processing, **21**-40–46
gate stack etch, **21**-21–33
origin of selectivity, **21**-43–44
oxide etch fundamentals, **21**-43–44
porous dielectrics, **21**-44–46
post etch metallization pre clean, **21**-46
profile control, **21**-26
shallow trench isolation, **21**-19–21
substrate contact dielectric etch, **21**-33–39
CMOS, multi gate, **11**-97–100
CMP. *See* Chemical mechanical polishing.

Co silicide process, **10**-12–18
Coatings, uniformity control and, **11**-71–74
Cobalt silicide, **11**-93–94
Cold wall CVD reactors, **13**-17
Collimated sputtering, **15**-13–15
Combined radiation shield, **11**-57–60
Commodities, semiconductor manufacturing and, **35**-12–14
Compensation control methods, **23**-20–24
  benefit payback, **23**-20–21
  controller goals, **23**-21–22
  feedback/feedforward, **23**-22
  real time, **23**-33–34
  run to run, **23**-22–33
Compensation process control, **23**-4–5
Complementary MOS devices (CMOS), **1**-1
Complex phase sequence, **10**-23–25
Complexing, **5**-5–6, **5**-7–8
Composition, **13**-13–14
Conducting CVD films, **13**-59–82
  ALD metal, **13**-80–82
  ALD transition, **13**-76–77
  aluminum, **13**-78–80
  copper, **13**-80
  dichlorosilane/tungsten hexafluoride, **13**-70–71
  monsilane/tungsten hexafluoride, **13**-68–69
  refractory metal nitride, **13**-76–77
  titanium, **13**-71–73
    nitride, **13**-73–76
  tungsten, **13**-59–67
    silicides, **13**-68
Conductive films, **15**-10
Confidence intervals, **34**-17–19
Confidence limit calculation, **22**-6–7
Congestion, **34**-3–4
Constrained systems, **23**-28–29
Contact electromigration, **30**-15–16
Contact etch stop layers, **21**-33
Contact feature challenge, **21**-54
Contact formation
  nickel silicide, **11**-95–96
  silicides, other types, **11**-96
  titanium nitride, **11**-96
Contact printing, **20**-2–3
Contact resistance, **21**-34–35
Contact resistivitiy, **10**-35–36
Contacts formation, **11**-92–97
  annealing, **11**-92
  cobalt silicide, **11**-93–94
  silicide, **11**-92
Contacts, self aligned, **21**-36–39
Contamination, metallic, **5**-16–17
Continuous Czochralski silicon growth, **3**-42–45
Controlled oxygen silicon crystal growth, **3**-24–28
Controlled vacancy concentration, **3**-66–67
  denuded oxygen, **3**-66–67
  oxygen precipitation, **3**-66–67
Controller system advanced process control, monitoring of, **23**-34–35
Controller, compensation control methods and, **23**-21–22
Convection flows, **3**-10–11
Coordinate systems, IP metrology and, **20**-55–56
Copper annealing, **11**-97
Copper CMP, **17**-45–46
  chemical effects and, **17**-40–42
Copper dual damascene interconnect, **5**-23
  first integration scheme, **5**-24
Copper electroplating
  damascene, **16**-1–42
  fundamentals of, **16**-8–13
    kinetics, **16**-8–12
      geometry effects, **16**-13
    mass transfer, **16**-12–13
Copper void detection, **24**-45–47
Copper, **13**-80
Copper/low-*k* dual damascene interconnect, associated defects, **5**-23–28
Copper/low-*k* integration, **2**-16–19
Corrosion, **30**-16–18
Cosmic ray neutrons, **31**-4–6, **31**-9–10
  boron, **31**-9–10
Cost of ownership, photomask and, **20**-69
Critical dimension bias, **21**-32–33
Critical dimension measurement, **24**-5–17
  application of, **24**-14–16
  calibration, **24**-4–17
    CD-SEM, **24**-5–10
    CD-SEM, focus, **24**-10–12
    scatterometry, **24**-12–14
    tools, **24**-5–10
  electrical, **24**-16–17
  overlay process control, **24**-17–20
Critical dimension scanning electron microscope. *See* CD-SEM.
Critical ionization model, **19**-23
Cross section analysis, **29**-15–16
Crucible dissolution, **3**-13–15
Crystal diameter, evolution of, **3**-38–42
Crystal orientations, **4**-18–20, **9**-12
  crystalline planes, **4**-19–20
  (100) plane rotation, **4**-18–19
Crystal rotations, **3**-15–16
Crystal structures, nickel silicide and, **10**-19
Crystalline planes, **4**-19–20
  hybrid orientation technology, **4**-19
CTE anisotropy, **10**-32
Cu CMP, random defects and, **27**-6
CVD. Chemical vapor deposition, **13**-1–82
  ALD process characterization, **13**-9–13
  atomic layer deposition, **13**-4
  basics of, **13**-2–16
  chemistry, **13**-2–5
    cleaning, **13**-4–5
    common dispositions, **13**-3
  definition of, **13**-1–2
  deposition kinetics, **13**-8–9
  film structure and properties, **13**-13–16
  high density plasma, **13**-24–24
  low pressure, **13**-25
  plasma enhanced, **13**-23–24
  reaction mechanisms, **13**-5–7
  reactors, **13**-17–21

ALD design, **13**-21–22
cold wall systems, **13**-17
hot wall systems, **13**-17
spin coating *vs.*, **13**-36–37
system design, **13**-16–22
reactor types, **13**-17–21
thin films, **13**-22–82
conducting, **13**-59–82
dielectrics, **13**-22–59
ALD high κ materials, **13**-52–55
barium strontium titanate, **13**-47–52
high κ, **13**-42–46
low κ dielectrics, **13**-35–42
polysilicon, **13**-55–56
process conditions, **13**-56–57
process parameter effect, **13**-57–59
silicon dioxide, **13**-22–31
silicon nitride, **13**-31–33
silicon oxynitride, **13**-33–35
uses of, **13**-1–2
Cycle time, **34**-2–3
Cyclic fatigue, **30**-21–23
CZ. *See* Czochralski technique.
Czochralski (CZ) technique, **3**-1
Czochralski melt, convection flows, **3**-10–11
Czochralski silicon growth, **3**-3–5
characteristics of, **3**-5–38
grown-in microdefects, **3**-32–38
impurity incorporation, **3**-7–32
thermal characteristics, **3**-5–7
continuous, **3**-42–45
controlled oxygen silicon crystal growth, **3**-24–28
magnetic field applied, **3**-28–32
microscopic inhomogeneity, **3**-18–24
nitrogen doping and, **3**-68
normal, **3**-24–28
oxygen incorporation
mechanism, **3**-13–18
segregation and, **3**-13–32
oxygen segregation, **3**-18–24

D

Damage, etching process modules and, **21**-39–40
Damascene copper electroplating, **16**-1–42
alternatives to, **16**-3
applications, **16**-4
chemistry, **16**-14–19
anode film, **16**-18–19
electrolytes, **16**-14–15
organic additives, **16**-15–18
deposition, **16**-4–8, **16**-19–39
alternate electrolytes, fill behavior, **16**-26
anneal behavior, **16**-34–36
defects, **16**-37–39
deposit planarity, **16**-28–29
electromigration, **16**-37
feature fill, **16**-19–23
field effects, **16**-31
fill evolution behavior, **16**-23–24
filling
additive behavior, **16**-28
current waveform, **16**-26
mass transfer impact, **16**-26
grain size, **16**-34–36
impurity levels of, **16**-32
leveling after fill, **16**-25–26
mass transfer, **16**-31–32
metallurgy, **16**-32–39
plating chemistry, **16**-24–25
properties of, **16**-32
reliability, **16**-32–39
resistivity, **16**-36–37
stress migration, **16**-37
surface roughness, **16**-28–29
terminal effect, **16**-30–31
thickness distribution, **16**-29
history of, **16**-4, 5
modeling capabilities, **16**-39–40
process, **16**-1–8
process control approaches, **16**-41–42
process integration, **16**-40–41
reasons for, **16**-2–3
Damascene copper integration, **2**-6–15
subtractive integration, comparison of, **2**-7–16
Damascene multi level metalization, random defects and, **27**-6
Damascene processing, **17**-3–4
Damascene structures, **17**-34–36
dishing, **17**-34–36
erosion, **17**-34–36
DARC. Dielectric anti reflection coating, **13**-33–35
Darkfield inspection, **27**-22
Data collection software, **25**-53–54
DC deceleration, **7**-50
Deal–Grove silicon oxidation model, **9**-6–8
Deceleration
DC type, **7**-50
ion implantation equipment and, **7**-49–58
Deep level transient spectroscopy (DLTS), **28**-19–21
Defect concentrations, **8**-4–5
kinds of defects, **8**-5
low level, **8**-4–85
Defect detection and analysis, yield management and, **27**-16–25
Defect structures, **3**-37–38
Defects, **16**-37–39
drying, **5**-17
electrical characterization techniques and, **28**-19–23
ion implantation and, **7**-24–30
physical and stress related, **28**-54–62
pole figures, **28**-60
Raman spectroscopy, **28**-60–62
X-ray
diffraction, **28**-54–57
reflectance, **28**-57–59
rocking curves, **28**-60
topography, **28**-57–59
surface preparation and, **5**-28–30
thermodynamics and, **8**-5–6

Delivery, photomask manufacturability and, **20**-68–69
Density, low-*k* dielectrics and, **12**-7
Denuded oxygen, **3**-66–67
Depletion transition, **4**-32
Deposit planarity, **16**-28–29
Deposition kinetics, **13**-8–9
Deposition process, **13**-43–44
Deposition, **9**-28–29
  directional, **15**-11–16
  ionized, **15**-16–23
  surface mobility based, **15**-10–11
Deprocessing, **29**-15
Deprotection kinetics, **19**-42
Design mitigation techniques, **31**-19
Design-in reliability, **22**-15–18
Detector, **25**-19–20
Develop step, **19**-37–40
  molecular weight, **19**-39–40
Developer, **19**-15–19
Development mechanisms of Novolac based photoresists, **19**-19–23
  critical ionization model, **19**-23
  membrane model, **19**-21
  percolation model, **19**-22–23
  secondary structure model, **19**-21
Device processing, **3**-64–66
DI technology, **4**-17–18
DI/ozone process, **5**-20–21
Diazosulfonyl compounds, **19**-28
Dichlorosilane/tungsten hexafluoride, **13**-70–71
Die exposure techniques, **29**-6–8
Dielectric anti reflection coating (DARC), **13**-33–35
Dielectric breakdown, time dependent, **30**-23–30
Dielectric CMP applications, **17**-42–44
Dielectric CVD, silicon dioxide and, **13**-25–26
Dielectric film metrology, **9**-14–23
  bulk trap, **9**-21–22
  common types, **9**-18
  film thickness measurement techniques, **9**-15–20
  gate oxide integrity, **9**-20–21
  high-*k* gate dielectrics, **9**-27–32
  interface trap measurements, **9**-21–22
  oxide charge, **9**-21–22
  ultra thin film measurements, **9**-22–23
Dielectric formation
  oxide improvement, **11**-80–82
  RTP and, **11**-80–83
  RTP applications and, **11**-76
    nitridation, **11**-76–79
    rapid thermal oxidation, **11**-76–79
  silicon nitridation, **11**-82
  steam, **11**-84–85
  trench features, **11**-82–83
Dielectric optical measurement, interconnect film thickness and, **24**-45
Dielectric processing, pre-metal, **11**-97
Dielectrics
  ALD applications and, **14**-28–32
  CVD thin films and, **13**-22–59
    ALD high κ materials, **13**-52–55
    barium strontium titanate, **13**-47–52
    high κ, **13**-42–46
    low κ dielectrics, **13**-35–42
    polysilicon, **13**-55–56
    process conditions, **13**-56–57
    process parameter effect, **13**-57–59
    silicon
      dioxide, **13**-22–31
      nitride, **13**-31–33
      oxynitride, **13**-33–35
  ellipsometric measurement and, **24**-25–27
  optical properties of, **11**-14
  post CMP cleaning, **17**-46–48
Different charge states, equilibrium and, **8**-5–6
Differential equation methods, **32**-12
Diffraction, X-ray, **28**-54–57
Diffuse reflectance spectroscopy, **25**-6–9
  band edge
    measurement techniques, **25**-7
    thermometry limitations, **25**-7–8
  temperature measurement, **25**-8–9
  theory, **25**-6–7
Diffusion controlled formation mechanism, **1**-26–27
Diffusion studies, **12**-18–20
Diffusion, random defects and, **27**-4
Digital control, ALD and, **14**-16–17
Dilute hydrofluoric acid, **5**-15
Dipole magnet mass resolving systems, **7**-46–49
Direct couple thermocouple, **11**-57–60
Direct thermocouple control, **11**-56–57
Directional deposition, **15**-11–16
  long throw, **15**-11–13
Dishing, **17**-34–36
Dislocation free growth, **3**-5–7
Dissolution mechanisms of Novolac based photoresists, **19**-9–19
  developer, **19**-15–19
  resin, **19**-10–12
  sensitizer, **19**-12–14
Distortion, lens, **18**-12
DLTS. *See* deep level transient spectroscopy.
DLVO theory, **5**-10–12
Dominantly diffusing species, **10**-27–28
  Kirkendall voiding, **10**-27–28
  reduction of bridging, **10**-27–28
Dopant activation, **11**-85–92
  limitations of conventional RTP, **11**-86–91
Dopant diffusion, **8**-1–17
  defects, thermodynamics of, **8**-5–6
  equilibrium formulation, **8**-9–12
    extrinsic diffusion, **8**-10–12
    intrinsic diffusion, **8**-9–10
  Fick's Laws, **8**-7–9
  non-equilibrium formulation, **8**-12–14
  point defects, **8**-3–4
    diffusion, **8**-6–7
    migration, **8**-6–7
  strained silicon, **8**-14–16
Dopants, **28**-39–46
  axial, **3**-7–8

effect on reactive phase formation, **10**-25–26
optical spectroscopy, **28**-43–45
secondary ion mass spectrometry, **28**-39–43
unintended, **3**-7–8
Doping process control, **24**-33–35
four point probe, **24**-33–35
junction depth measurement via carrier illumination, **24**-35
optically modulated optical reflection, **24**-34
secondary ion mass spectrometry, **24**-34–35
Dose control, **7**-52–58
in situ implant, **7**-57–58
Double gate SOI MOSFETS, **4**-40–41
Drain contract issues, **1**-36–42
DRAM applications, **13**-42
Dry chrome etching, **20**-43–44
inductive coupled plasma, **20**-44
ion milling, **20**-44
magnetic enhanced RIE, **20**-44
plasma
applications, **20**-44
reactors, **20**-44
reactive ion etching, **20**-44
Drying defects, surface preparation and, **5**-17
Drying, supercritical carbon dioxide cleaning processes and, **6**-17–18
Dual beam scanning system, **7**-55–57
Dual damascene copper integration, **2**-6–15
challenges of, **2**-10
Dual inlaid processing, **21**-40–46
Dual metal gate, **21**-52
Dynamic modeling, **34**-16–21
Monte Carlo discrete event simulation, **34**-16–19
queuing models, **34**-19–21
Dynamic temperature uniformity control, **11**-70–71

E

e-Beam writer, **20**-19–27
beams and imaging impact, **20**-22–23
resist sensitivity, **20**-25–27
scanning mechanism, **20**-20–22
raster scan system, **20**-20–21
stage movement, **20**-22
vector scan system, **20**–22
source and optics, **20**-19–20
E-model, **30**-24–25
Economic effects, semiconductor manufacturing and, **35**-4–8
Economic models, semiconductor manufacturing and, **35**-9
Economics of semiconductor manufacturing, **35**-1–20
Edge effect, CMP and, **17**-28–29
Edge rounding, **3**-47
Effective channel length, **28**-14–15
Effective segregation coefficient k_{eff}, **3**-9–12
Elastic constraint effects, **12**-9–12
Electrical CD measurement, **24**-16–17
Electrical characterization techniques, **28**-1–30
capacitance voltage profiling, **28**-7–9
charge based measurements, **28**-23–25
deep level transient spectroscopy, **28**-19–21
defects and carrier lifetimes, **28**-19–23
failure site isolation and, **29**-4–5
tools for, **29**-5–6
four point probe, **28**-2–6
interface states, **28**-17–18
lateral doping profiling, **28**-9–12
mobile charge, **28**-17–18
modulated photoreflectance, **28**-6–7
MOSFET, **28**-12–17
oxide integrity, **28**-18–19
probe measurements, **28**-25–30
resistivity and carrier concentration, **28**-2–12
SOI materials and devices, **4**-26–28
Electrical methods, film thickness measurement techniques and, **9**-18
Electrical modeling
methodologies, **32**-12–13
differential equation methods, **32**-12
integral equation methods, **32**-12–13
parameters, **32**-10–12
capacitance, **32**-10–11
inductance, **32**-12
resistance, **32**-10–11
Electrochemistry, **5**-2–4
oxidation, **5**-3
redox, **5**-2
standard half-cell potentials, **5**-3
Electrodeposition, **16**-4–8
Electrolytes, **16**-14–15
fill behavior, **16**-26
Electromagnetic beam scanning system, **7**-52–55
Electromigration, **16**-37, **30**-11–15
contact, **30**-15–16
Electron beam inspection, **27**-22–24
Electron beam probing, **29**-11–12
Electronic stopping, **7**-2–8
channeling, **7**-5–6
random, **7**-3–5
Electrophoretic effects, **5**-10–12
Electrostatic beam scanning system, **7**-52–55
Elevated source/drain, **3**-55–56
Ellipsometric measurement
alternate dielectrics, **24**-25–27
gate dielectic film thickness and, **24**-21–27
multi wavelength, **24**-23
poly Si thickness, **24**-27
spectroscopic, **24**-23
theory, **24**-21–23
thin gate films, **24**-24–25
ultra thin SiO_2, **24**-25–27
Ellipsometric systems, **24**-24
Ellipsometry, **9**-15, **25**-48–49
ELTRAN, **4**-14
Emission microscopy, **29**-9
Emissivity effects, pyrometry and, **11**-38–39, **11**-41–45
End station, **7**-59–61
Endpoint detection, **25**-14–15
evolving window factor analysis, **25**-16–17
neural network, **25**-15–16

Endpoint methods, **17**-13–14
Energy contamination, **7**-48
Enthalpy of formation, **8**-4
Environment, slurry and, **17**-22–23
Environmental effects, low-*k* dielectrics and, **12**-15–20
Environmental health and safety, plasma etching and, **21**-17
Epi resistivity and thickness, **25**-50–53
Epitaxial growth, **3**-49–61
  heteroepitaxy, **3**-51–52
  selective, **3**-52–57
  silicon epitaxial wafer, **3**-49–61
  strained silicon, **3**-57–61
Epitaxy, solid phase, **11**-91
Equilibrium formulation
  dopant diffusion and, **8**-9–12
  extrinsic diffusion, **8**-10–12
  intrinsic diffusion, **8**-9–10
Equilibrium segregation coefficient, **3**-22
Equipment
  definition of, **22**-2
  failure, definition, **22**-2
  intended functions, **22**-2–3
  operating environment, **22**-3
  reliability, **22**-1–26
    definition of, **22**-2
    metrics, **22**-3–24
    repairable system, **22**-3
Equipment reliability
  high level performance metrics, **22**-8–11
  improvement of, **22**-12–19
    build-in reliability, **22**-18–19
    design-in reliability, **22**-15–18
    goal allocation, **22**-13–14
    goals and requirements, **22**-13–14
    managing growth, **22**-19
  maintainability metrics, **22**-8
  SEMI E10, **22**-25–27
  testing of, **22**-20–24
    life cycle phases, **22**-23–24
    types, **22**-21–22
Equipment reliability discipline, use in business practices, **22**-24–25
Equipment signal monitoring, **23**-19–20
Error budgets, **18**-36–38
Errors, pyrometry and, **11**-38–39
Etching deposition, random defects and, **27**-5–6
Etching interactions, **21**-47–48, **21**-59
Etching process modules, CMOS, **21**-18–46
Etching stop, aspect ratio, **21**-35–36
Etching
  aspect ratio dependent, **21**-15
  back end of line, **21**-40–46
  chemical, **21**-8–9
  endpoint detection, **25**-29
  gate, **21**-52
    soft landing, **21**-27
    stack, **21**-21–33
  high-κ dielectric, **21**-52–53
  main gate electrode, **21**-25–27
  nitride, **5**-17–18
  oxide, **5**-15
  oxide/nitride, **5**-18
  polysilicon, **5**-17
  shallow trench isolation, **21**-19–21
  spacer, **21**-54
  substrate contact dielectric, **21**-33–39
  supercritical carbon dioxide cleaning processes and, **6**-15–17
  wet chemical, **5**-15
Evolving window factor analysis, **25**-16–17
EWMA. See Exponentially weighted moving average.
EWMA based controllers, **23**-27–28
Exhaust gas monitoring applications, **25**-17–18
Exponentially weighted moving average. *See* EWMA.
Exposure bake kinetics, **19**-34–37
Exposure bake step, **19**-31–37
Exposure step, **19**-25–42
  chemistry, **19**-26–29
    diazosulfonyl compounds, **19**-28
    nitrobenzyl sulfonate esters, **19**-28–29
    onium salts, **19**-26
    sulfonyloxy imides, **19**-26–27
  develop, **19**-37–40
  kinetics, **19**-29–31
Exposure time ALD process characterization, **13**-10
Exposure tool systems, **18**-10–19
  alignment and overlay, **18**-10–17
  throughput, **18**-17–19
Extension implants, **7**-34–36
Extracting materials properties, **12**-4–8
Extraction optics, **7**-42–46
  Bernas sources, **7**-42–45
Extrinsic diffusion, **8**-10–12
Extrinsic point defects, **8**-3

F

Fab design, **34**-5–9
Factories, semiconductor manufacturing and, **35**-18–20
Factory modeling, **34**-1–21
  dynamic, **34**-16–21
  fab design, **34**-5–9
  performance metrics, **34**-1–5
    congestion, **34**-3–4
    cycle time, **34**-2–3
    inventory, **34**-3
    Little's Law, **34**-3
    performance levels, **34**-4
    throughput, **34**-1–2
  static, **34**-10–16
Failure analysis, **29**-1–22
  chemical characterization, **29**-18–21
    Fourier transform infrared spectroscopy, **29**-20
    secondary ion mass spectroscopy, **29**-20
    X-ray analysis, **29**-19
  definitions of, **29**-2–3
  future of, **29**-21
  physical tools for, **29**-14–18

cross-section, **29**-15–16
deprocessing, **29**-15
microscopy, **29**-16–17
package analysis, **29**-14
parallel polishing, **29**-15
site isolation, **29**-4–14
Failure mechanisms, **30**-11–15
Failure site isolation, **29**-4–14
die exposure techniques, **29**-6–8
electrical characterization, **29**-4–5
tools for, **29**-5–6
global techniques, **29**-8–10
probing, **29**-10–14
Failure statistics, **30**-10
False negatives, **23**-16–18
False positives, **23**-16–18
Fault detection, process control abnormalities and, **23**-19
Fault detection and classification, **25**-55–56
Fault interdiction, **25**-56
Fault prognosis, **25**-56
FBE. *See* floating body effects.
FDC analysis, **25**-53–54
Feature fill, **16**-19–23
mechanism, **16**-19–23
Feature profile, **21**-15–16
Feedback/feedforward control, **23**-22
Fick's laws of diffusion, **8**-7–9
Field effect, **16**-31
Fill behavior, electrolytes and, **16**-26
Fill evolution behavior, **16**-23–24
Fill performance
additive behavior, **16**-28
current waveform, **16**-26
mass transfer impact, **16**-26
Fill response, plating chemistry and, **16**-24–25
Fill, leveling after, **16**-25–26
Film flatness, CMP control and, **24**-47–49
Film structure
composition, **13**-13–14
CVD and, **13**-13–16
interface properties, **13**-16
microstructures, **13**-14–15
step coverage, **13**-15–16
stress, **13**-15
Film thickness effects, **12**-3–4
Film thickness measurement techniques, **9**-15–20
electrical methods, **9**-18
ellipsometry, **9**-15
high resolution transmission electron microscopy, **9**-15, **9**-18
misc. methods, **9**-20
optical interference, **9**-15
Films
ARC, **15**-10
conductive, **15**-10
planar, **15**-10
Fixed wavelength systems, **25**-12
Flip chip packages
modeling verification, **32**-24–26
thermal deformation of, **32**-18–22
Float zone silicon growth, **3**-2–3
Czochralski silicon growth, **3**-3–5
neutron transmutation doping (NTD), **3**-2
Floating body effects (FBE), **4**-28
Flow fill low κ dielectric film, **13**-38
Fluorine doped silicate glass, **13**-37–38
Focused ion beam
repair, **20**-60–61
sample preparation, **28**-35–37
systems, **24**-49
Formation mechanism, **10**-26–29
diffusion controlled, **1**-26–27
dominant diffusing species, **10**-27–28
rapid diffusion, disadvantages of, **10**-28–29
45 32 nm technology nodes
advanced patterning schemes, **21**-47–48
aspect ratio, **21**-48–50
challenges for, **21**-50–51
materials, **21**-50
silicon structures, **21**-51
transistor structures, **21**-50–51
etch interactions, **21**-47–48
extension of, **21**-60–61
imaging resolution, **21**-50
immersion optical lithography, **21**-46–47
line edge roughness, **21**-50
line etch process
back end, **21**-55–57
front end, **21**-52–55
plasma etching and, **46**-57
Four gate FET, **4**-44–45
Four point probe, **24**-33–35, **28**-2–6
Fourier transform infrared, **25**-17–18
exhaust gas monitoring applications, **25**-17–18
spectroscopy, **29**-20
Microspot techniques, **29**-20
theory, **25**-17
Front end line etch processes, **21**-52–55
Full wafer exposure, **18**-18
Fully depleted SOI MOSFETS, **4**-29–33
depletion transition, **4**-32
meta-stable dip, **4**-31
subthreshold slope, **4**-30
threshold voltage, **4**-29–30
transconductance, **4**-30–31
volume inversion, **4**-32
Furnaces, **9**-2–5

G

Gas phase reactant concentration, **25**-10–25
acoustic composition measurement, **25**-22–26
Fourier transform infrared, **25**-17–18
mass spectroscopy/residual gas analysis, **25**-18–22
optical emission spectroscopy, **25**-11–16
Gate all around SOI MOSFETS, **4**-43–44
Gate dielectric film thickness,
ellipsometric measurement, **24**-21–27
alternate dielectrics, **24**-25–27
multi wavelength, **24**-23

poly Si thickness, **24**-27
spectroscopic, **24**-23
thin gate films, **24**-24–25
ultra thin SiO_2, **24**-25–27
ellipsometric systems, **24**-24
Gate dielectric technology, **9**-1–33
Gate dielectric thickness, measuring of, **24**-32–33
Gate dielectrics, **9**-23–32
high-*k*, **9**-27–32
plasma nitride oxide, **9**-23–27
Gate dimension thinning, **21**-23–25
Gate doping, **11**-91–92
Gate etch, **21**-52
Gate fabrication integration issues, **21**-27–32
chamber walls, **21**-27–30
microtrenching profile generation, **21**-30
uniformity, **21**-30–32
Gate induced floating body effect (GIFBE), **4**-37
Gate oxide integrity (GOI), **7**-65–68, **9**-20–21
hot carrier degradation, **9**-21
oxide breakdown, **9**-20–21
Gate oxide thickness, electrical measurement, **24**-27–33
Gate stack etch, **21**-21–33
critical dimension bias, **21**-32–33
gate dimension thinning, **21**-23–25
gate fabrication integration issues, **21**-27–32
hard mask trim, **21**-23–25
main gate electrode etch, **21**-25–27
overetch, **21**-27
photoresist, **21**-23–25
soft landing, **21**-27
Gate stack issues, **1**-22–28
Gate, dual metal, **21**-52
Gaussian profiles, **7**-8–9
Ge on insulator, **4**-23–24
Ge, introduction of, **10**-16
Generation lifetimes, **28**-22–23
Generic mask on wafer, **20**-2–5
contact printing, **20**-2–3
projection printing, **20**-3–4
proximity printing, **20**-2–3
reduction-projection printing, **20**-4–5
GIFBE. *See* gate induced floating body effect.
Glass, **20**-40
Global planarization, **17**-3
Global techniques, **29**-8–10
carrier generation, **29**-10
charging based, **29**-9
emission microscopy, **29**-9
hot spot detection, **29**-8–9
thermal generation, **29**-9–10
GOG mask technology, **20**-7
alternating phase shift mask, **20**-7–13
GOI. *See* gate oxide integrity.
Grain size, damascene copper electroplating and, **16**-34–36
Grinding, **3**-47, **17**-2
Grown in microdefects, **3**-32–38
defect structures, **3**-37–38
relevance, **3**-32–36

H

Half-cell potentials, standard, **5**-3
Halftone phase shift mask, **20**-13–14
Halide hayride chemical compounds, **14**-8–9
Halogens, **9**-5
Hard mask trim etch, **21**-23–25
Harmonic signature analysis, **25**-29–30
Heat sink attachments, **32**-8–9
Heat transfer models, **11**-34–38
Heating, wafer thermal response and, **11**-28–32
Helium, hydrogen co-implantation and, **4**-12–13
Heteroepitaxy, **3**-51–52
Hexafluoride, **13**-68–71
HF-last, initiation process and, **14**-17–18
HfO_2, high-*k* gate dielectrics and, **9**-29–32
HfSiON, high-*k* gate dielectrics and, **9**-29–32
High density plasma CVD, **13**-23–24
High energy cosmic ray neutrons, **31**-4–6
High perveance beams, **7**-48–49
High refractive index polymers, **19**-50
High resistivity silicon, **3**-68–70
High resolution transmission electron microscopy, **9**-15, **9**-18
High spatial resolution imaging, **28**-30–39
focused ion beam sample preparation, **28**-35–37
scanning
electron microscopy, **28**-30–32
probe microscopy, **28**-37–39
transmission electron microscopy, **28**-32–35
High temperature annealing, **4**-15
High voltage SOI devices, **4**-5
High κ dielectric, **13**-42–46
deposition process, **13**-43–44
DRAM applications, **13**-42
precursor effect on electrical properties, **13**-45–46
step coverage, **13**-44
Ta precursors, **13**-42–43
Ta_2O_5 DRAM applications, **13**-42
Ta_2O_5 integration, **13**-44
RTP applications and, **11**-83
High-κ dielectric etch, **21**-52–53
High-κ gate dielectrics, **9**-27–32
deposition, **9**-28–29
HfO_2, **9**-29–32
HfSiON, **9**-29–32
High/low pressure effects, oxidation and, **9**-11–12
Higher-κ oxide capacitors, **14**-25–32
History effects, **4**-28–29
Hot carrier degradation, **9**-21
Hot carriers, **28**-15–17
Hot plates, **11**-51–53
Hot spot detection, **29**-8–9
Hot wall CVD reactors, **13**-17
HOT. *See* hybrid orientation technology.
HT. PSM fabrication, **20**-14–15
Hybrid orientation technology (HOT), **4**-19
Hydrofluoric acid, **5**-15–16
metallic contamination, **5**-16–17
Hydrogen helium co-implantation and, **4**-12–13

Hydrogen implanted silicon wafers, annealed, **4**-11–12
Hydrogen induced splitting, **4**-10
  annealed hydrogen-implanted silicon wafers, **4**-11–12
  implanted silicon wafers, **4**-11
Hygro thermo mechanical behavior, **32**-13–27
  analysis of, **32**-15–18
  flip chip packages, thermal deformation, **32**-18–22
  material properties, **32**-23
  solder joint fatigue failure, **32**-15
Hysteresis, latch and, **4**-28

I

I/E model, **30**-25
ILD planarization, **17**-43–44
Image evaluation, **20**-65–67
  aerial analysis, **20**-66–67
  system description, **20**-65–66
Imaging, **18**-4–7
Imaging limitation, **21**-59
Imaging resolution, **21**-50
Imaging theory, **20**-6–7
Immersion fluids, **19**-47–50
Immersion lithography, **18**-38–43
  defects, **18**-42
  projection lens optical design, **18**-40–41
  thermal control, **18**-43
  topcoats for, **19**-47
  water resist surface effects, **18**-42–43
Immersion optical lithography, **21**-46–47
Implant angle integrity, **7**-74–77
Implant optimization, **4**-16
Implant, random defects and, **27**-4
Implanted silicon wafers, non-annealing, **4**-11
Implanter architecture drivers, **7**-42
Impurities, **28**-39–46
  damascene copper electroplating and, **16**-32
  radioactive decay of, **31**-6–8
  trace element radiochemical methods, **28**-45–46
Impurity incorporation, **3**-7–32
  axial dopant distribution, **3**-7–8
  Czochralski silicon growth, **3**-13–32
  macroscopic inhomogeneity, **3**-9–12
  microscopic inhomogeneity, **3**-9–12
  unintended dopants, **3**-8–9
In line metrology, **24**-1–50
  front end processes, **24**-20–36
    doping process control, **24**-33–35
    gate dielectric film thickness, **24**-21–27
    gate oxide thickness, **24**-27–33
    measurement techniques, **24**-32–33
      gate dielectric thickness, **24**-32–33
      nitrogen concentration, **24**-32–33
    stress enhanced carrier mobility, **24**-35–36
  in FAB FIB, **24**-49
  interconnect process control, **24**-36–49
  lithography processes, **24**-5–19
  manufacturing sensitivity analysis, **24**-4–5
  measurement precision to process tolerance ratio, **24**-2–4
In situ contamination monitoring, **27**-12–13
In situ implant dose control, **7**-57–58
In situ metrology, **25**-1–56
  potential sensor measurement techniques, **25**-47–53
    ellipsometry, **25**-48–49
    epi resistivity and thickness, **25**-50–53
  process state sensors, **25**-4–31
  software, **25**-53–54
    data collection, **25**-53–54
    FDC analysis, **25**-53–54
    model based process control, **25**-54
  use in semiconductor manufacturing, **25**-55–56
    fault
      detection and classification, **25**-55–56
      interdiction, **25**-56
      prognosis, **25**-56
    model based process control, **25**-56
  wafer state sensors, **25**-31–47
Incorporation of oxygen, **3**-13–32
Inductance, **32**-12
Inductive coupled plasma, **20**-44
Inductively coupled plasmas, **21**-11–12
Inhomogeneity, macroscopic and microscopic, **3**-9–12
Initiation processes
  ALD and, **14**-17–18
  HF-last, **14**-17–18
  metal nitrides, **14**-18
  nanolaminate interfaces, **14**-18
Inspection of mask, **20**-58–59
Inspection, SOI wafers and, **4**-26
Integral equation methods, **32**-12–13
Integrated circuit packaging, **32**-1–27
  challenges of, **32**-2–9
  electrical modeling, **32**-9–13
  hygro-thermo-mechanical behavior, **32**-13–27
  thermal management, **32**-8–9
    heat sink attachments, **32**-8–9
Integrated circuits, terrestrial radiation, **31**-1–21
Integrated emissivities, **11**-18–20
Integrated network modeling, **34**-21
Integrated optical properties, **11**-7
Integrated processing, RTP applications and, **11**-83
Integration
  dual damascene copper integration, **2**-6–15
  interconnect-copper, **2**-1–15
  issues, **21**-17
  low-*k*, **2**-16–19
Interbay transport and storage, **33**-10–24
  major elements, **33**-11–17
  system planning, **33**-17–24
Interconnect, **14**-3, **14**-31–32
  copper dual damascene, **5**-23
  copper integration, **2**-1–15
  subtractive aluminum, **5**-21–22
Interconnect applications, PVD and, **15**-9–23
Interconnect film thickness, **24**-37–47
  acoustic measurement methods, **24**-39–44
  dielectric optical measurement, **24**-45
  metal film thickness, **24**-45

metal illumination, **24**-45–47
porous low *k*, **24**-47
x-ray methods, **24**-37–39
Interconnect process control, **24**-36–49
chemical mechanical polishing process control, **24**-47–49
film thickness, **24**-37–47
Interconnect structures formation, **11**-96–97
copper annealing and reflow, **11**-97
curing low κ films, **11**-97
pre metal dielectric processing, **11**-97
Interface properties, **13**-16
Interface states, **28**-17–18
Interface trap measurements, **9**-21–22
Interfacial adhesion, **12**-15–16
Interfacial boundary layer, **3**-9–12
Interferometric technique, **25**-36–40
International Technology Roadmap for Semiconductors (ITRS), **1**-9
Interstitial point defects, **8**-3–4
Intrabay systems, **33**-24–49
automated, **33**-31–34
benefits of, **33**-47–50
configurations, **33**-30–31
implementation barriers, **33**-35
performance factors, **33**-35–40
sizing and planning, **33**-41–46
support interfaces, **33**-46
throughput constraints, **33**-40
Intrinsic diffusion, **8**-9–10
Intrinsic point defects, **8**-3
Inventory, **34**-3
Ion implantation, **7**-1–80
applications of, **7**-32–40
buried layers, **7**-32–33
source/drain extension implants, **7**-34–36
channeling, **7**-20–24
damage annealing, **11**-85–92
rapid thermal annealing, **11**-85–86
defects, **7**-24–30
amorphization, **7**-28–30
standard modeling, **7**-25–28
electronic stopping, **7**-2–8
equipment, **7**-40–61
architecture drivers, **7**-42
beam scanning system, **7**-52–58
dose control, **7**-52–58
end station, **7**-59–61
extraction optics, **7**-42–46
general requirements of, **7**-41–42
ion sources, **7**-42–46
post analysis acceleration, and deceleration, **7**-49–58
wafer charging control, **7**-58
mass analysis systems, **7**-46–49
nuclear stopping, **7**-2–8
overview, **7**-2
physics and material science of, **7**-2–32
process control, **7**-61–80
beam-residual gas interactions control, **7**-61–65
charging, **7**-65–68
gate oxide integrity, **7**-65–68
heavily implanted photoresist removal, **7**-71–74
implant angle integrity, **7**-74–77
metrology, **7**-79–80
photoresist mask integrity, **7**-68–71
wafer contamination, **7**-77–79
wafer cooling, **7**-68–71
source/drain implants, **7**-33–34
statistics, **7**-8–20
Gaussian profiles, **7**-8–9
Monte Carlo approaches, **7**-13, 17, 19–20
Pearson IV distribution, **7**-9–12
symbol definitions, **7**-30–32
Ion mass spectrometry, **28**-39–43, **29**-20
Ion milling, **20**-44
Ion sources, **7**-42–46
molecular, **7**-45–46
Ionized deposition, **15**-16–23
applications, **15**-21–23
Ionized plasma, **21**-6–8
Ionizer, **25**-19
Ions in matter, **31**-2–3
IP metrology, **20**-53–55
calibration, **20**-55
coordinate systems, **20**-55–56
introducing of, **20**-53–55
Isolated dense bias, **21**-14–15
Isolation, **1**-42, **21**-54
ITOX process, **4**-16–17
ITRS (International Technology Roadmap for Semiconductors), **1**-9

J

Jobdeck creation, **20**-33–34
Junction depth measurement via carrier illumination, **24**-35

K

Kinetics, **16**-8–12, **19**-29–31
exposure bake, **19**-34–37
geometry effects, **16**-13
Kink effect, **4**-28
Kirchhoff's Law, **11**-6
Kirkendall voiding, **10**-27–28

L

Lapping, **3**-47, **17**-2
Large diameter silicon growth, **3**-38–45
continuous Czochralski, **3**-42–45
crystal diameter, **3**-38–42
Laser repair, **20**-60
Laser writer, **20**-27–31
raster scan, **20**-27–28
SLM programmable mask, **20**-28–31
Latch, hysteresis and, **4**-28
Lateral doping profiling, **28**-9–12

Learning curve, semiconductor manufacturing and, **35**-9–12
Lens
  distortion, **18**-12
  projection, **18**-40–41
Leveling after fill, **16**-25–26
Life cycle phase testing, equipment reliability and, **22**-23–24
Limitations, optical lithography manufacturing and, **18**-32
Limited reactions, rapid ALD and, **14**-19–20
Limited yield analysis, **26**-9–11
Lindhard–Scharff–Schiott regime, **7**-4
Line edge roughness, **19**-43, **21**-50
Line etch process
  back end, **21**-55–57
    low-κ, **21**-56–57
    metal line resistivity, **21**-56
    porous ULK dielectrics, **21**-55–56
    ultra low-κ dielectrics, **21**-55–57
  front end, **21**-52–55
    contact feature challenges, **21**-54
    dual metal gate, **21**-52
    gate etch, **21**-52
    high-κ dielectric etch, **21**-52–53
    isolation, **21**-54
    spacer etch, **21**-54
    STI, **21**-54
Linearity with cycling, ALD and, **14**-16–17
Lithographic processes, **19**-2–5, **24**-5–9
  critical dimension measurement, **24**-5–17
Local planarization, **17**-3
Logic design, photomask device specific reticle issues and, **20**-71
Lognormal distribution, **30**-7
Long throw directional deposition, **15**-11–13
  applications of, **15**-15–16
  collimated sputtering, **15**-13–15
Look aheads, **23**-14
Low dose SIMOX, **4**-16
Low energy cosmic ray neutrons, **31**-9–10
Low pressure CVD, **13**-25
Low temperature oxide, **13**-24–25
Low *k* air, **21**-56–57
Low *k* dielectrics, **12**-1–20, **13**-35–42
  channel crack failures, **12**-3–8
  chemical composition, **12**-7
  CVD vs. spin coating, **13**-36–37
  density, **12**-7
  elastic constraint effects, **12**-9–12
  environmental effects, **12**-15–20
    channel cracking, **12**-16–20
    interfacial adhesion, **12**-15–16
  mechanical properties, **12**-7
  pattern layout effects, **12**-13–15
  silicon based films, **13**-37–42
Low *k* films curing, **11**-97
Low *k* integration, **2**-1–19
Low/high pressure effects, oxidation and, **9**-11–12

M

Macro inspection, **27**-24
Macroscopic
  inhomogeneity, **3**-9–12
    radial impurity uniformity, **3**-11–12
  microscopic inhomogeneity and, **3**-9–12
    Czochralski melt, **3**-10–11
    effective segregation coefficient k_{eff}, **3**-9–12
    interfacial boundary layer, **3**-9–12
  radial impurity uniformity, **3**-11–12
Magnet mass resolving systems
  ribbon beams, **7**-48
  species and energy contamination, **7**-48
Magnetic enhanced RIE (MERIE), **20**-44
Magnetic field applied Czochralski growth (MCZ), **3**-28–32
Magnification, **18**-12
Main gate electrode etch, **21**-25–27
Maintainability metrics, **22**-8
Manufacturing sensitivity analysis, **24**-4–5
Manufacturing, optics for, **18**-7–10
Market dynamics, economics of semiconductor manufacturing and, **35**-1–2
Market pricing, semiconductor manufacturing and **35**-6–8
Mask cycle time, **20**-68
Mask data creation, **20**-32–33
Mask error factor, **18**-34–36
Mask errors, **18**-12
Mask inspection, **20**-58–59
Mask interactions, **21**-16–17
Mask writers, **20**-19–39
  data preparation, **20**-31–32
    jobdeck creation, **20**-33–34
    machine type relationship, **20**-35–39
    mask data creation, **20**-32–33
    runtimes, **20**-39
  e-Beam, **20**-19–27
  laser, **20**-27–31
Mask yield, **20**-68
Mass analysis beams, high perveance beams, **7**-48–49
Mass analysis systems, **7**-46–49
  dipole magnet resolving systems, **7**-46–49
Mass filter, **25**-19–20
Mass transfer, **16**-12–13, 31–32
  impact, **16**-26
Material handling systems, **33**-50–53
  anomaly handling, **33**-54–55
  applications, **33**-50
  computer simulation, **33**-55–57
  configuration, **33**-52
  elements, functions and requirements, **33**-51
  implementation of, **33**-57–64
  performance factors, **33**-53
  reliability of, **33**-54
Material properties, **32**-23
MCZ. *See* magnetic field applied Czochralski growth.
Measurement lag, **23**-31

Measurement precision to process tolerance ratio, resolution *vs*., **24**-2–4
Mechanical probing, **29**-10–11
Mechanical properties, low-*k* dielectrics and, **127**-7
Megasonic cleaning, **5**-12–15
Membrane model, **19**-21
Memory, photomask device specific reticle issues and, **20**-70
Memory SER sensitivity, **31**-13–15
MEMS, **4**-5
MERIE. *See* magnetic enhanced RIE.
Metal, ALD, **13**-80–82
Metal CMP applications, **17**-44–46
  aluminum, **17**-46
  copper, **17**-45–46
  tungsten, **17**-45
Metal film thickness, interconnect film thickness and, **24**-45
Metal gates, **10**-38–43, **11**-97–100
  ALD applications and, **14**-28–32
  fully silicided gates, **10**-39–40
Metal hydroxide
  solubility, **5**-6
  values, **5**-6
Metal illumination
  copper void detection, **24**-45–47
  metal line thickness, **24**-45–47
Metal line
  resistivity, **21**-56
  thickness, **24**-45–47
Metal nitrides, **14**-18
Metal organic chemistry, **14**-9–10
Metal organic TiN films, **13**-74–75
Metal oxide, **13**-52–55
Metal oxide semiconductor (MOS), **1**-1
Metallic contamination, **5**-16–17
Metallurgy, **16**-32–39
Metals
  initiation processes and, **14**-18
  optical properties of, **11**-14, 16
  post CMP cleaning, **17**-46–48
Meta-stable dip, **4**-31
Metrics calculations, example of, **22**-11–12
Metrology, **20**-47–55, **23**-14, **7**-79–80
  CD, **20**-48–53
  dielectric film of, **9**-14–23
  errors, overlay, **18**-14
  in line, **24**-1–50
  in situ, **25**-1–56
  IP, **20**-53–55
Microdefects, grown-in, **3**-32–38
Microprocessors, **4**-5
  photomask device specific reticle issues and, **20**-70
Microscopic inhomogeneity, **3**-9–12, **3**-18–24
  automatic diameter control-induced perturbation, **3**-12
  equilibrium segregation coefficient, **3**-22
  non-centrosymmetric thermal distribution, **3**-12
  oxygen precipitation, **3**-23–24
  thermal convection-related temperature fluctuations, **3**-12
Microscopy, **29**-16–17
Microsensors, **4**-5
Microspot Fourier transform infrared spectroscopy, **29**-20
Microstructural effects, barium strontium titanate and, **13**-51–52
Microstructures, **13**-14–15
Microtrenching profile generation, **21**-30
Millisecond annealing, **11**-88–90
Mitigation techniques, terrestrial radiation, **31**-16–21
Mobile charge, **28**-17–18
Mobile ions/surface inversion, **30**-27–28
Mobility issues, **4**-36–37
Model based process control, **25**-56
  software, **25**-54
Modeling, ion implantation defects and, **7**-25–28
Modified illumination, **18**-23–26
Modulated photoreflectance, **28**-6–7
Molecular ion sources, **7**-45–46
Molecular weight, **19**-39–40
Molybdenum silicide, **20**-40–41
Monochromators, **25**-12
Monsilane/tungsten hexafluoride, **13**-68–69
Monte Carlo approaches, **7**-13, 17, 19–20
Monte Carlo discrete event simulation, **34**-16–19
  confidence intervals, **34**-17–19
  probability distributions, **34**-17–19
  run times, **34**-16–17
  simulation mechanics, **34**-16–17
  verification and validation, **34**-19
Moore's Law, **35**-2–4
Morphological stability, **10**-29–30, 36–38
MOS devices (metal oxide semiconductor), **1**-1
  advanced concepts, **1**-44–53
    multiple gate, **1**-44–49
    other semiconductor types, **1**-51–53
    SOI substrates and devices, **1**-44
    summary of, **1**-52
    transport enhanced, **1**-50–51
  characteristics of, **1**-3–8
MOSFET (silicon metal oxide semiconductor field effect transistor), **1**-1
MOSFET device characterization, **28**-12–17
  effective channel length, **28**-14–15
  hot carriers, **28**-15–17
  source drain resistance, **28**-14–15
  threshold voltage, **28**-12–14
MOSFET device scaling, **1**-8–21
  performance of, **1**-12–21
  rules, **1**-9–12
MOSFET manufacturing, **1**-22–44
  channel doping issues, **1**-28–36
  gate stack issues, **1**-22–28
  isolation, **1**-42
  source/drain contract issues, **1**-36–42
  substrate, **1**-42
  thermal budget issues, **1**-42–44
Multi wavelength ellipsometric, **24**-23
Multi-gate CMOS, **11**-97–100
Multiple gate
  MOS devices, **1**-44–49
    concepts, **1**-49

SOI MOSFETS, **4**-40–45
double gate, **4**-40–41
four gate FET. **4**-44–45
gate all around, **4**-43–44
triple gate, **4**-41–43
Multivariable systems, **23**-28–29

N

N CAR negative based on cross linking, **20**-42
N CAR positive based on chain scission, **20**-41–42
N CAR positive based on dissolution inhibition, **20**-42
N Car. *See also* on chemically amplified resists.
n^+ silicon growth crystal, **3**-17–18
n^+, oxygen precipitation and, **3**-63–64
Nanoelectric SOI devices, **4**-5–6
Nanolaminate interfaces, **14**-18
Nanolaminates, **14**-19
Narrow channels, **4**-34
Near atmospheric processing, **9**-5–11
Network decomposition, **34**-20–21
Neural network endpoint detection, **25**-15–16
Neutron transmutation doping (NTD), **3**-2
Neutrons, cosmic ray, **31**-4–6, 9–10
Nickel silicide, **10**-18–38, **11**-95–96
anomalous thermal expansion, **10**-31–32
unit cell dimensions, **10**-31
basic properties of, **10**-18–21
crystal structures, **10**-19
phase diagram, **10**-18–19
volumetric change, **10**-19
contact resistivity, **10**-35–36
formation mechanism, **10**-26–29
integrity, **10**-29–31
formation of $NiSi_2$, **10**-29
morphological stability, **10**-29–30
NiSi films, **10**-30–31
morphological stability, **10**-36–38
phase stability, **10**-35–36
properties of, **10**-19–21
properties of, silicon consumption, **10**-21
reactive phase formation, **10**-21–26
$Si_{1-x}Ge_x$ devices, **10**-34–35
texture development, **10**-33–34
NiSi films, nickel silicide and integrity, **10**-30–31
$NiSi_2$ formation, nickel silicides and integrity, **10**-29
Nitric based silicon oxynitride films, **9**-13
Nitridation, **11**-76–79
Nitride etch, **5**-17–18
Nitrobenzyl sulfonate esters, **19**-28–29
Nitrogen, **9**-23–25
Nitrogen concentration, measuring of, **24**-32–33
Nitrogen doping, CZ silicon and, **3**-68
Nitrous based silicon oxynitride films, **9**-13
Non centrosymmetric thermal distribution, **3**-12
Non contact sensors, RTP temperature measurement and, **11**-45, 46
Non linear systems, **23**-28–29
Non selective epitaxial growth, **3**-55
SiGE HBT, **3**-55
Non-equilibrium formulation, **8**-12–14
Non-patterned wafers, **17**-25–29
Novel BOX, **4**-39–40
Novolac based photoresists, **19**-5–23
development mechanisms, **19**-19–23
dissolution mechanism, **19**-9–19
overview, **19**-5–9
NTD. *See* neutron transmutation doping.
Nuclear stopping, **7**-2–8
Nucleation process, **13**-62–67
ALD W, **13**-63–67
Numerical modeling, oxidation and, **9**-13–14

O

OES. *See* optical emission spectroscopy.
Ohmic contacts, **21**-33
On chemically amplified resists (n-CAR), **20**-41–42
n-CAR negative based on cross linking, **20**-42
n-CAR positive based on
chain scission, **20**-41–42
dissolution inhibition, **20**-42
On insulator substrates, **4**-23–24
(100) plane rotation, **4**-18–19
Onium salts, **19**-26
OPC (Optical proximity correction), **20**-15–19
Open loop lamp intensity control (OLIC), **11**-45, 47–51
Optical CD metrology, **20**-48–49
Optical emission spectroscopy (OES), **25**-11–16
applications, **25**-14–16
end point detection, **25**-14–15
calibration and maintenance, **25**-14
charge coupled device array spectrometers, **25**-12–14
fixed wavelength systems, **25**-12
monochromators, **25**-12
spectrographs, **25**-12
Optical interference, **9**-15
Optical lithography, **18**-1–47
exposure tool system considerations, **18**-10–19
immersion, **18**-38–43, **21**-46–47
manufacturing considerations, **18**-32–38
aberrations, **18**-7–10
error budgets, **18**-36–38
limits of, **18**-32
mask error factor, **18**-34–36
mix and match concerns, **18**-36
process latitude, **18**-32–33
patterning basics, **18**-2–7
generators, **18**-2–3
imaging basics, **18**-4–7
replicators, **18**-3–4
patterning roadmaps, **18**-45–46
progress in, **18**-38–45
polarization, **18**-43–45
resolution enhancement techniques, **18**-19–31
modified illumination, **18**-23–26
optical proximity effect, **18**-27–31
phase shift masks, **18**-20–23, 25

Optical masks, **2**-6–19
advanced technology, GOG, **20**-7
phase shift masks, **20**-7
HT.PSM fabrication, **20**-14–15
imaging theory, **20**-6–7
OPC, **20**-15–19
SRAF, **20**-15–19
Optical properties
dielectrics, **11**-14
metals and silicides, **11**-14, 16
silicon, **11**-10–14
wafers and, **11**-10–15
Optical proximity correction, **20**-15–19
Optical proximity effect, **18**-27–31
correction computation, **18**-30–31
correction techniques, **18**-29–30
Optical sensors, **25**-32–40
interferometric technique, **25**-36–40
reflectometry technique, **25**-32–36
Optical spectroscopy, **28**-43–45
Optically modulated optical reflection, **24**-34
Optics for manufacturing, **18**-7–10
Optics, e-beam writer and, **20**-19–20
Organic additives, **16**-15–18
Origin of selectivity, **21**-43–44
Overall equipment efficiency, **22**-9–10
Overetch, **21**-27
Overlay process control, CD measurement and, **24**-17–20
Overlay
analysis of, **18**-14–17
lens distortion, **18**-12
magnification, **18**-12
mask errors, **18**-12
metrology errors, **18**-14
lithography and, **18**-10–12
stage errors, **18**-12–13
thermal processing wafer distortion, **18**-13
wafer alignment keys, **18**-13–14
wafer chucking errors, **18**-14
Oxidation interactions, **9**-12–13
crystal orientation, **9**-12
segregation, **9**-12
silicon oxynitride films, **9**-12–13
Oxidation state, **5**-7
Oxidation technology, **9**-1–33
annealing ambient, **9**-5
dielectric films, metrology of, **9**-14–23
furnaces, **9**-2–5
halogens, **9**-5
numerical modeling, **9**-13–14
current models, **9**-13–14
limitations of, **9**-14
rapid thermal processors, **9**-2–5
theory of, **9**-5–12
Oxidation theory, near atmospheric processing, **9**-5–11
Deal–Grove silicon oxidation model, **9**-6–8
high/low pressure effects, **9**-11–12
thin silicon model, **9**-8–11
Oxidation, **5**-3
salicide development and, **10**-14–15
Oxide breakdown, gate oxide integrity and, **9**-20–21
Oxide charge, **9**-21–22
Oxide dielectric CMP, chemical effects and, **17**-36–37
Oxide etching, **5**-15
buffered hydrofluoric acid, **5**-15–16
dilute hydrofluoric acid, **5**-15
fundamentals, **21**-43–44
Oxide integrity, **28**-18–19
Oxide/nitride etch, **5**-18
Oxides
dielectric formation and, **11**-80–82
random defects and, **27**-5–6
Oxidizer/sulfuric chemistry, **5**-18–19
Oxygen behavior
controlled vacancy concentration, **3**-66–67
oxygen precipitation kinetics, **3**-61–67
Oxygen incorporation mechanism, **3**-13–32
crucible dissolution, **3**-13–15
crystal rotations, **3**-15–16
n^+ silicon growth crystal, **3**-17–18
p^+ silicon growth crystal, **3**-17–18
surface evaporation, **3**-13–15
Oxygen precipitation, **3**-23–24, 66–67
Oxygen precipitation kinetics, **3**-61–67
device processing, **3**-64–66
n^+, **3**-63–64
p^+, **3**-63–64
Oxygen segregation, **3**-13–32
Oxygen silicon crystal growth, controlled, **3**-24–28

P

p^+ silicon growth crystal, **3**-17–18
p^+, oxygen precipitation and, **3**-63–64
Package analysis, **29**-14
Pad condition design, **17**-11–12
Pad mechanics, **17**-29–31
Parallel polishing, **29**-15
Parasitic bipolar transistor (PBT), **4**-28
Partially depleted SOI MOSFETS, **4**-28–29
floating body effects, **4**-28
history effects, **4**-28–29
hysteresis and latch, **4**-28
kink effect, **4**-28
parasitic bipolar transistor, **4**-28
transient effects, **4**-28–29
Partially ionized plasma, **21**-6–8
Particle adhesion, **5**-8–9
Particle removal, **5**-8
chemical undercutting, **5**-9–10
DLVO theory, **5**-10–12
electrophoretic effects, **5**-10–12
megasonics, **5**-12–15
Particulates, cleaning of, **6**-11–14
Parylenes, **13**-40–41
Pattern effect, uniformity control and, **11**-72–74
Pattern generators, **18**-2–3
Pattern layout effects, **12**-13–15
Pattern replicators, **18**-3–4
Pattern transfer, chrome etching, **20**-43–46
Patterned buried oxide, **4**-17

Patterned wafers, **17**-29–36
Patterning basics, optical lithography and, **18**-2–7
Patterning schemes, **21**-47–48
PBT. *See* parasitic bipolar transistor.
Pearson IV distribution, **7**-9–12, 14–18
Pellicles, **20**-46–47
Percolation model, **19**-22–23
Performance metrics, **22**-8–11
  availability, **22**-9
  hierarchy, **22**-11
  overall equipment efficiency, **22**-9–10
  utilization, **22**-9
Perturbation, automatic diameter control, **3**-12
PH, metal hydroxide solubility, **5**-6
Phase diagram, nickel silicides and, **10**-18–19
Phase measuring, **20**-63–65
Phase shift masks (PSM), **18**-20–23, 25, **20**-7
Phase stability, **10**-35–36
Phosphorus doped silicon glass, **13**-30–31
Photo acoustic metrology, **25**-41–44
  applications, **25**-42–44
  hardware configuration, **25**-42
  measurement technique, **25**-41–42
Photolithography, **5**-18, **27**-4–5
Photomask device specific reticle issues, **20**-69–71
  analog design, **20**-71
  ASIC design, **20**-70
  logic design, **20**-71
  memory specific, **20**-70
  microprocessor specific, **20**-70
Photomask fabrication, **20**-1–71
  chrome, **20**-40
  glass, **20**-40
  material homogeneity, **20**-41
  materials and processing, **20**-39–47
  molybdenum silicide, **20**-40–41
  pellicles, **20**-46–47
  photoresist, **20**-41–42
  writing patterns, **20**-19–39
Photomask processing, **20**-42–46
  pattern transfer, **20**-43–46
  resist coat and develop, **20**-43
Photomask qualification, **20**-47–67
  defects and repair of, **20**-59–63
    AFM nanomatching mask repair, **20**-62–63
    focused ion beam repair, **20**-60–61
    laser repair, **20**-60
  defects, **20**-56–59
    types of, **20**-56–58
  image evaluation, **20**-65–67
  inspection and repair, **20**-56–63
  mask inspection, **20**-58–59
  metrology, **20**-47–55
  phase and transmission measurement, **20**-63–65
Photomask, cost of ownership, **20**-69
Photomask, device specific reticle issues, **20**-69–71
Photomask, evolution in usage, **20**-2–19
  generic mask on wafer, **20**-2–5
  optical masks, **2**-6–19
Photomask, manufacturability, **20**-67–71
  cost, **20**-69
  mask cycle time, **20**-68
  on-time delivery performance, **20**-68–69
  manufacturability, yield, **20**-68
Photomask, structure, **20**-2–19
Photoresist, **20**-41–42, **21**-23–25
  ArF materials, **19**-40–46
  chemically amplified resists (CAR), **20**-42
  high refractive polymers, **19**-50
  immersion lithography, topcoats for, **19**-47
  mask integrity, **7**-68–71
  materials, **19**-1–53
  on chemically amplified resists, **20**-41–42
  post-ArF material requirements, **19**-50–53
  processing, **19**-1–53
  relief image formation, **19**-1–5
  novolac based, **19**-5–23
  removal, **7**-71–74
  strip, **5**-18
  striping and removal, **6**-6–10
Phototonics, SOI materials and devices and, **4**-45
Physical analysis tools
  failure and, **29**-14–18
  transmission electron microscopy, **29**-17–18
Physical characterizations, **28**-30–62
Physical constants, **A**-1
Physical vapor deposition (PVD), **15**-1–25
  interconnect applications, **15**-9–23
    ARC films, **15**-10
    conductive films, **15**-10
    directional deposition, **15**-11–16
    ionized deposition, **15**-16–23
    planar films, **15**-10
    reflow, **15**-10–11
    surface mobility-based deposition, **15**-10–11
  semiconductor applications, **15**-1–2
  sputtering, **15**-2–6
  systems, **15**-6–9
Physics
  plasma and, **21**-4–10
  process uniformity control and, **11**-63–65
Pilots, **23**-13
Piranha. *See* sulfuric peroxide.
Planar films, **15**-10
Planarization, **17**-2
  global, **17**-3
  local, **17**-3
Plasma applications, **20**-44
Plasma assisted ALD, **14**-10–11
Plasma chemistry, **21**-4–10
  partially ionized, **21**-6–8
  surface, **21**-8–10
Plasma enhanced CVD, **13**-23–24
Plasma etching, **21**-1–64
  CMOS etch process modules, **21**-18–46
  45-32 nm technology nodes, **21**-46–57
  plasmas relationship to, **21**-2–17
  22 nm nanotechnology, future of, **21**-57–63
Plasma etching issues, **21**-13–17
  aspect ratio dependent etching, **21**-15
  charging, **21**-17
  environmental health and safety, **21**-17

feature profile, **21**-15
integration issues, **21**-17
isolated-dense bias, **21**-14–15
loading, **21**-14–15
mask interactions, **21**-16–17
selectivity, **21**-14
sidewall shape, **21**-15–16
uniformity, **21**-13–14
Plasma etching processes, modeling of, **21**-63–64
Plasma generation, **21**-4
Plasma nitrided oxide, **9**-23–27
nitrogen, **9**-23–25
reliability of, **9**-27
SiON scaling, **9**-27
Plasma physics, **21**-4–10
Plasma power measurements to, **25**-30
Plasma, properties of, **21**-4
Plasma reactors, **20**-44
Plasma sheath, **21**-5
Plasma treatment cycle, **13**-75–76
Plasmas, **21**-2–17
etching tools, **21**-10–13
capacitively coupled, **21**-10–11
inductively couple, **21**-11–12
other types, **21**-12–13
Platen, **17**-2
Plating chemistry, fill response to, **16**-24–25
Plug formation, **13**-62
PMOS recessed source/drain, **3**-56–57
Point beam, **20**-22–23
Point defects, **8**-3–4
diffusion, **8**-6–7
extrinsic, **8**-3
interstitial, **8**-3–4
intrinsic, **8**-3
migration, **8**-6–7
vacancy, **8**-3
Polarization, **18**-43–45
Pole figures, **28**-60
Polish tool integration, **17**-12–13
Polishing patterned wafers, **17**-29–36
damascene structures, **17**-34–36
pad mechanics, **17**-29–31
spatial averaging, **17**-31–34
Polishing, **3**-48, **17**-2
Poly Si thickness, **24**-27
Polymer deposition, **21**-9
Polymers, high refractive index, **19**-50
Polysilicon, **13**-55–56
CMP applications, **17**-44
etch, **5**-17
Pore size distribution, porous low *k* and, **24**-47
Porous dielectrics, etching on, **21**-44–46
Porous low *k*, pore size distribution, **24**-47
Porous ULK dielectrics, **21**-55–56
Post analysis acceleration and deceleration, **7**-49–58
Post ArF material requirements, **19**-51–53
Post CMP cleaning, **17**-12–13
Potential sensor measurement techniques, **25**-47–53
ellipsometry, **25**-48–49
epi resistivity and thickness, **25**-50–53
Power measurements to plasma, **25**-30
Pre clean etching process, **21**-46
Pre metal dielectric processing, **11**-97
Precision, interconnect film thickness and, **24**-44–45
Precursors
barium strontium titanate and, **13**-47–51
effect on electrical properties, **13**-45–46
Predictor corrector control, **23**-30
Probability distributions, **34**-17–19
Probe measurements, **28**-25–30
Probing, **29**-10–14
backside, **29**-13–14
component isolation, **29**-12–13
electron beam, **29**-11–12
mechanical, **29**-10–11
Process capability indices, **23**-14–16
Process conditions, **13**-56–57
Process control, **16**-41–42
abnormality detection, **23**-18–20
equipment signals, **23**-19–20
fault detection, **23**-19
other methods, **23**-19–20
sensitivity of, **23**-20
univariate statistical, **23**-18–19
advanced, **23**-5–7
basic concepts, **23**-14–18
error types, **23**-16–18
process capability indices, **23**-14–16
process qualifications, **23**-14
compensation methods, **23**-20–34
continuous improvement, **23**-35–36
controller system advanced process control, monitoring of, **23**-34–35
error types
false negatives, **23**-16–18
false positives, **23**-16–18
interconnect, **24**-36–49
ion implantation, **7**-61–80
model based, **25**-56
overview of, **23**-1–37
reducing effective system noise, **23**-36
run to run controller, monitoring of, **23**-34–35
semiconductor manufacturing
benefits of, **23**-9–14
history of, **23**-7–9
look-aheads, **23**-14
metrology, **23**-14
operational practices, **23**-14
pilots, **23**-13
variation
requiring compensation, **23**-10–11
time scale, **23**-11–13
software, **25**-54
systematic yield loss, **23**-2–4
types, **23**-4–7
abnormality, **23**-4
compensation, **23**-4–5
Process integration, **16**-40–41
feature profile, **16**-40–41
seed interaction, **16**-40–41

Process model, **23**-25–26
benefits of, **23**-26
issues with, **23**-26–27
Process qualifications, **23**-14
Process state sensors, **25**-4–31
gas phase reactant concentration, **25**-10–25
RF properties, **25**-26–30
temperature, **25**-4–10
acoustic wafer temperature sensor, **25**-9–10
diffuse reflectance spectroscopy, **25**-6–9
pyrometry, **25**-5–6
wall deposition, **25**-30–31
Process technology mitigation techniques, **31**-17–18
Process uniformity control, **11**-60, 63–74
non temperature problems, **11**-63
physics of, **11**-63–65
configurations, **11**-65–69
dynamic temperature, **11**-70–71
optimization procedures, **11**-69–70
steady-state non-uniformity, **11**-63–64
temperature differences, **11**-65
transient, **11**-65
Process window, **21**-36
Processing latitude, optical lithography manufacturing and, **18**-32–33
Product sensitivity analysis, **26**-11–16
Projection lens design, **18**-40–41
Projection printing, **20**-3–4
Properties, CVD and, **13**-13–16
Proximity printing, **20**-2–3
PSM. *See* phase shift masks.
Pulsed laser annealing, **11**-90–91
Purge time optimization, **13**-10
PVD. *See* Physical vapor deposition.
Pyrometry, **25**-5–6
calibration of, **11**-41
emissivity effects, **11**-38–39
errors in, **11**-38–39
wafer emissivity effects, **11**-41–45
wavelength choice, **11**-40–41

Q

Queuing models, **34**-19–21
applications of, **34**-19–20
definition of, **34**-19–20
integrated network modeling, **34**-21
network decomposition, **34**-20–21

R

Radial impurity uniformity, **3**-11–12
Radiation shields, **11**-53–56
combined, **11**-57–60
Radio frequency linear acceleration, **7**-50–51
Radioactive decay
alpha particles, **31**-6–8
impurities and, **31**-6–8
Radiochemical methods, trace elements and, **28**-45–46
Raman spectroscopy, **28**-60–62
stress measurements and, **4**-26
Random defects
sources and types, **27**-2–7
chemical mechanical polishing, **27**-5–6
Cu CMP, **27**-6
damascene multi level metalization, **27**-6
diffusion and implant, **27**-4
etch deposition, **27**-5–6
oxides, **27**-5–6
photolithography, **27**-4–5
surface preparation, **27**-4
wafer edge engineering, **27**-7
yield
limits, **26**-7–9
models, **26**-4–7
Random stopping, **7**-3–5
Bethe–Bloch regime, **7**-3
interpolation between regimes, **7**-5
Lindhard–Scharff–Schiott regime, **7**-4
Rapid ALD, limited reactions, **14**-19–20
Rapid diffusion, formation mechanism and, **10**-28–29
Rapid growth cycles, **35**-5–6
Rapid thermal annealing (RTA), **11**-85–86
Rapid thermal oxidation (RTO)
kinetics of, **11**-76–79
steam and, **11**-84
Rapid thermal processing. *See* RTP.
Raster scan system, **20**-20–21, 27–28
RBS. *See* Rutherford backscattering spectrometry.
RCA cleaning process, **5**-2
Reaction mechanism, **5**-19, **13**-5–7
ALD, **13**-7–8
thermally activated reaction, **13**-5–7
Reactive ion etching (RIE), **20**-44, **21**-9
Reactive phase formation, **10**-21–26
complex phase sequence, **10**-23–25
dopants effect on, **10**-25–26
thermal budget, **10**-21
Reactors, CVD, **13**-17–21
Real time compensation control methods, **23**-33–34
Recessed source/drain, PMOS, **3**-56–57
Recombination lifetimes, **28**-21–22
Redox, **5**-2
Reduction-projection printing, 204–5
Reflow, **11**-97, **15**-10–11
Refractive index polymers, **19**-50
Refractory metal nitride, **13**-76–77
Relaxed SiGe substrate, strained silicon epitaxy and, **3**-58–61
Relectrometry technique, **25**-32–36
Reliability
accelerated testing, **30**-2–3
engineering, **30**-1–30
equipment and, **22**-2–6
metrics
applications, **22**-4–6
analytical/theoretical values, **22**-5
observed values, **22**-6
values, **22**-5

calculations, **22**-4–5
confidence limit calculation, **22**-6–7
use of, **22**-7–8
physics, **30**-1–30
time to failure
modeling, **30**-3–6
statistics, **30**-6–10
Relief image formation, **19**-1–5
chemically amplified, **19**-23–40
lithographic process, **19**-2–5
Repair techniques, photomask qualification and, **20**-59–63
Residual gas analysis, **25**-18–22
conventional style, **25**-18–20
detector, **25**-19–20
ionizer, **25**-19
mass filter, **25**-19–20
sensor type, **25**-21–22
Resin, **19**-10–12
Resist coat and develop, **20**-43
Resist sensitivity, **20**-25–27
Resist stripping, **20**-45–46
Resistance, **32**-10–11
salicide development and, **10**-16–18
Resistivity, **16**-36–37
electrical characterization techniques and, **28**-2–12
Resolution
interconnect film thickness and, **24**-44–45
measurement precision to process tolerance ratio *vs.*, **24**-2–4
Reticle handling metrics, **33**-3
RF properties, **25**-26–30
applications, **25**-29–30
etching endpoint detection, **25**-29
harmonic signature analysis, **25**-29–30
power measurements, **25**-30
measurement technologies, **25**-27–28
sensor
installation, **25**-28
technologies, **25**-26–27
signal processing, **25**-28
Ribbon beams, **7**-48
RIE. *See* reactive ion etching.
Roadmaps for optical lithography requirements, **18**-45–46
Rocking curves, **28**-60
Root cause isolation, **27**-27–28
RTA. *See* rapid thermal annealing.
RTO. *See* rapid thermal oxidation.
RTP (Rapid thermal processing)
applications of, **11**-2–3
contacts formation, **11**-92–97
dielectric formation and processing, 11–76
high-κ, **11**-83
nitridation, **11**-76–79
rapid thermal oxidation, **11**-76–79
steam, **11**-84–85
emerging types, **11**-97–100
metal gates, **11**-97–100
multi gate CMOS, **11**-97–100
strain engineering, **11**-97–100
gate doping, **11**-91–92
integrated processing, **11**-83
interconnect structures, **11**-96–97
ion implantation damage annealing and dopant activation, **11**-85–92
RTP process uniformity control, **11**-60, 63–74
RTP systems
configurations, **11**-3–5
control technology, **11**-1–74
dynamics, **11**-33–34
hardware, **11**-1–74
modeling of, **11**-27–28, 29–30
semiconductor processing, **11**-74–100
temperature measurement, **11**-38–60
combined radiation shield, **11**-57–60
direct couple thermocouple, **11**-57–60
direct thermocouple control, **11**-56–57
hot plates, **11**-51–53
non-contact sensors, **11**-45
open loop lamp intensity control, **11**-45, 47–51
problems, **11**-38
pyrometry, **11**-38–45
radiation shields, **11**-53–56
thermocouples, **11**-45, 47, 48
thermal radiation physics, **11**-5–7
wafers
optical properties, **11**-10–15
thermal response, **11**-27–38
RTP thin dielectric formation, **11**-80–83
oxide improvement, **11**-80–82
silicon, **11**-82
trench features, **11**-82–83
Rules of MOSFET device scaling, **1**-9–12
Run to run compensation control, **23**-22–33
additive disturbance assumption, **23**-29–30
algorithms, **23**-25
compare to other methods, **23**-30–31
constrained systems, **23**-28–29
EWMA based controllers, **23**-27–28
measurement lag, **23**-31
multivariable, **23**-28–29
non-linear, **23**-28–29
predictor corrector control, **23**-30
process model, **23**-25–26
Run to run controller, monitoring of, **23**-34–35
Rutherford backscattering spectrometry (RBS), **28**-46–49

S

Safety, slurry and, **17**-22–23
Salicide
comparisons of, **10**-6
development, **10**-5–18
Co silicide process, **10**-12–18
Ge, **10**-16
oxidation, **10**-14–15
resistance and, **10**-16–18
Si reduction and effect of, **10**-15–16
surface cleanings, **10**-14–15
Ti silicide process, **10**-7–12

Saturation characteristics, ALD and, **14**-13–15
SC-1, **5**-7
  dilution, effects of, **5**-14
Scaled devices, performance of, **1**-12–21
Scaling rules, **1**-9–12, 21
  International Technology Roadmap for Semiconductors, **1**-9
Scaling trends, **4**-33–40
  channel thickness, **4**-34–36
  mobility issues, **4**-36–37
  narrow channels, **4**-34
  novel BOX, **4**-39–40
  short channels, **4**-33–34
  ultra-thin gate dielectrics, **4**-37–39
Scanning electron microscopy, **28**-30–32
Scanning mechanism, **20**-20–22
Scanning probe microscopy, **28**-37–39
Scatterometer, **25**-44–47
Scatterometry, **24**-12–24
Schottky barrier source drain, **10**-38–43
  silicide, **10**-40–43
Secondary ion mass spectrometry (SIMS), **24**-34–35, **28**-29–43, **29**-20
Secondary structure model, **19**-21
Seed interaction, **16**-40–41
SEG. *See* selective epitaxial growth.
Segregation of oxygen, **3**-13–32
Segregation, oxidation interactions and, **9**-12
Selective epitaxial growth (SEG), **3**-52–57
  CMOS elevated source/drain, **3**-55–56
  non, **3**-55
  PMOS recessed source/drain, **3**-56–57
  SiGe HBT, **3**-53–55
Selective nitride etch, **5**-17–18
Self aligned
  contact, **21**-36–39
  silicide process, **10**-3–5
SEM CD metrology, **20**-49–52
SEMI E10 specifications, **22**-25–27
Semiconductor cleaning, supercritical carbon dioxide, **6**-1–21
Semiconductor devices
  complementary MOS, **1**-1
  introduction to, **1**-1–3
  metal oxide, **1**-1
  silicon metal oxide semiconductor field effect transistor (MOSFET), **1**-1
  terrestrial radiation effects on, **31**-10–13
  very large scale integration integrated circuits, **1**-1
Semiconductor manufacturing
  commoditized, **35**-12–14
  economic
    effects, **35**-4–8
      capacity, **35**-6–8
      market pricing, **35**-6–8
      rapid growth cycles, **35**-5–6
    models, **35**-9
  economics of, **35**-1–20
    factories and equipment, **35**-18–20
    learning curve, **35**-9–12
    market dynamics, **35**-1–2
    Moore's Law, **35**-2–4
    technological advances, **35**-14–18
  process control and, **23**-7–9
  standards, **C**-1–4
Semiconductor processing, RTP systems and, **11**-74–100
Semiconductor wafers, thermal radiative properties and, **11**-15–27
  absorptivities, **11**-15–27
  integrated emissivities, **11**-18–20
  silicon spectral emissivity, **11**-15–27
Semiconductors
  MOS device advanced concepts and, **1**-51–53
  PVD and, **15**-1–2
Sensitivity analysis, **24**-4–5
Sensitizer, **19**-12–14
Sensor type residual gas analysis, **25**-21–22
Sequential self-limiting processes, **14**-5–6
Sequential/combinational logic SER sensitivity, **31**-15–16
Shallow trench isolation etch, **21**-19–21
Shaped beam, **20**-22–23
Sheath, plasma and, **21**-5
Short channels, **4**-33–34
Si, BOX thickness measurements and, **4**-24–26
Si reduction, effect on salicide development, **10**-15–16
$Si_{1-x}Ge_x$ devices, **10**-34–35
Sidewall shape, **21**-15–16
SiGe HBT, **3**-53–55
SiGe substrate, relaxed, **3**-58–61
Silicon
  optical properties of, **11**-10–14
  spectral emissivity of, **11**-15–27
  22 nm nanotechnology and, **21**-62–63
Silcon coatings, **11**-20–26
Silicide formation, **11**-92
  titanium, **11**-92–93
Silicide gates, **10**-39–40
Silicide Schottky barrier source drain, **10**-40–43
Silicides, **10**-1–43
  history of, **10**-2–3
  metal gate, **10**-38–43
  nickel, **10**-18–38
  optical properties of, **11**-14, 16
  Schottky barrier source drain, **10**-38–43
  self-aligned process, **10**-3–5
Silicon based low κ dielectric film, **13**-37–42
  black diamond, **13**-39–40
  carbon based, **13**-40
  flow fill, **13**-38
  fluorine doped silicate glass, **13**-37–38
  future of, **13**-42
  parylenes, **13**-40–41
  silsequioxane, **13**-38
Silicon consumption, nickel silicides and, **10**-21
Silicon crystal growth processes, **3**-2–5
  float zone silicon growth, **3**-2–3
Silicon dioxide, **13**-22–31
  boron doped silicon glass, **13**-30–31
  borophosphosilicate glass, **13**-28–30
  dielectric CVD, **13**-25–26
  high density plasma CVD, **13**-24–24
  low pressure CVD, **13**-25

low temperature oxide, **13**-24–25
phosphorus doped silicon glass, **13**-30–31
plasma enhanced CVD, **13**-23–24
undoped silicon glass, **13**-26–28
Silicon epitaxial wafer, **3**-49–61
Silicon epitaxy, strained, **3**-57–61
Silicon glass
boron doped, **13**-30–31
undoped, **13**-26–28
Silicon growth
crystal, **3**-17–18
large diameter, **3**-38–45
Silicon materials, **3**-1–62
epitaxial growth, **3**-49–61
large diameter silicon growth, **3**-38–45
new types, **3**-68–70
high resistivity, **3**-68–70
nitrogen doping, **3**-68
processing and oxygen behavior, **3**-61–67
silicon crystal growth processes, **3**-2–5
ultra large scale integration ICs, **3**-1
wafer preparation, **3**-45–49
Silicon melt, non-centrosymmetric thermal distribution and, **3**-12
Silicon metal oxide semiconductor field effect transistor (MOSFET), **1**-1
Silicon nitridation, **11**-82
Silicon nitride, **13**-31–33
Silicon on insulators. *See* SOI.
Silicon oxynitride, **13**-33–35
dielectric anti reflection coating, **13**-33–35
films, **9**-12–13
ammonia-based, **9**-12–13
nitric based, **9**-13
nitrous, **9**-13
Silicon structure, **21**-51
Silsequioxane, **13**-38
SIMOX, **4**-4, **4**-14–18
concepts of, **4**-14–15
high temperature annealing, **4**-15
implant optimization, **4**-16
ITOX process, **4**-16–17
low-dose, **4**-16
patterned buried oxide, **4**-17
SIMS. *See* secondary ion mass spectrometry.
Simulation mechanics, **34**-16–17
Single wafer systems, **14**-23–24
SiON scaling, **9**-27
SiON, future as
dielectric replacement, **9**-32–33
gate electric, **9**-32
Slicing, **3**-45–46
SLM programmable mask, **20**-28–31
architecture, **20**-30
chip fabrication, **20**-30–31
Slurry, **17**-16–24
composition of, **17**-17–20
abrasives, **17**-17–18
chemistries, **17**-18–20
distribution of, **17**-20–21
environmental issues, **17**-22–23
handling issues, **17**-22–23
manufacturing issues, **17**-20–24
parameter measurements, **17**-21–22
quality monitoring, **17**-21
safety issues, **17**-22–23
Smart Cut™
SOI material manufacturing and, **4**-3
technology, **4**-8–13
hydrogen and helium co-implantation, **4**-12–13
hydrogen-induced splitting, **4**-10
process description, **4**-9–10
Soft landing gate etch, **21**-27
Software, in situ metrology and, **25**-53–54
SOI (Silicon on insulators).
advanced wafer engineering, **4**-18–24
basics of, **4**-3–6
manufacturing of, **4**-3–4
bond and etchback, **4**-4
ELTRAN, **4**-4
SIMOX, **4**-4
smart cut, **4**-3
electrical characterization, **4**-26–28
devices, **4**-27–28
wafers, **4**-26–27
formation methods, **4**-6
fully depleted SOI MOSFETS, **4**-29–33
materials and devices, **4**-1–46
multiple gate SOI MOSFETS, **4**-40–45
partially depleted SOI MOSFETS, **4**-28–29
photonic applications of, **4**-45
scaling trends, **4**-33–40
SOI wafers, physical characterization of, **4**-24–26
workings of, **4**-4–6
high voltage models, **4**-5
innovations in, **4**-5
MEMS, **4**-5
microprocessors, **4**-6
microsensors, **4**-5
nanoelectric types, **4**-5–6
SOI MOSFETS
fully depleted, **4**-29–33
multiple gate, **4**-40–45
partially depleted, **4**-28–29
SOI strained, **4**-20–23
SOI substrates and devices, **1**-44
SOI wafer fabrication methods, **4**-6–18
bond and etchback, **4**-13–14
ELTRAN, **4**-14
other types, **4**-17–18
DI technology, **4**-17–18
SON, **4**-18
SOS, **4**-18
SIMOX, **4**-14–18
Smart Cut™ technology, **4**-8–13
wafer bonding, **4**-7–8
SOI wafers, physical characterization of, **4**-24–26
inspection for particles and defects, **4**-26
Si and BOX thickness measurements, **4**-24–26
stress measurements, **4**-26
structural defects, **4**-25
surface roughness, **4**-25

Solder joint fatigue failure, **32**-15
Solid phase epitaxy, **11**-91
Solubility, **5**-5–6
  metal hydroxide values, **5**-6
Solvation capabilities, **6**-5–6
SON, **4**-18
Source drain resistance, **28**-14–15
Source/drain contract issues, **1**-36–42
Source/drain extension implants, **7**-34–36
  ultra-shallow junction formation, **7**-37–40
Source/drain implants, **7**-33–34
Space charge effect, **20**-25
Spacer etch, **21**-54
Spatial averaging, **17**-31–34
Species contamination, **7**-48
Spectral emissivity, **11**-6
  calculations of, **11**-8–10
  silicon and, **11**-15–27
Spectrographs, **25**-12
Spectrometry, ion mass, **28**-39–43
Spectroscopic ellipsometry, **24**-23
Spectroscopy
  auger electron, **28**-53–54
  deep level transient, **28**-19–21
  diffuse reflectance, **25**-6–9
  ion mass, **29**-20
  Raman, **28**-60–62
  X-ray photoelectron, **28**-51–53
Spike annealing, **11**-87–88
Spin coating, CVD *vs.*, **13**-36–37
Sputtering, **15**-2–6, **21**-8
  collimated, **15**-13–15
SRAF (sub resolution assist features), **20**-15–19
Stage errors, **18**-12–13
Standards for semiconductor manufacturing, **C**-1–4
Static financial projections, **34**-13–16
Static modeling, **34**-10–16
  static financial projections, **34**-13–16
Steady state non uniformity, **11**-63–64
Steam, rapid thermal oxidation, **11**-84
Step coverage, **13**-15–16, 44
STI, **21**-54
Stop layers, contact etching, **21**-33
Stopping, electronic and nuclear, **7**-2–8
Strain engineering, **11**-97–100
Strained silicon epitaxy, **3**-57–61
  relaxed SiGe substrate, **3**-58–61
Strained silicon on insulator, **4**-20–23
Strained silicon, diffusion and, **8**-14–16
Stress defects, **28**-54–62
Stress enhanced carrier mobility, measurement of, **24**-35–36
Stress measurements, Raman spectroscopy and, **4**-26
Stress migration, **16**-37, **30**-18–21
Stress, **13**-15
Striping and removal, photoresist, **6**-6–10
Structural defects, SOI wafers and, **4**-25
Sub resolution assist features. *See* SRAF.
Substrate contact dielectric etch, **21**-33–39
  aspect ratio etch stop, **21**-35–36
  contact resistance, **21**-34–35
  ohmic contacts, **21**-33
  process window, **21**-36
  self aligned contact and selectivity, **21**-36–39
  stop layers, **21**-33
Substrate temperature, effect of, **21**-10
Substrate, **1**-42
Substrates, on-insulator, **4**-23–24
Subthreshold slope, **4**-30
Subtractive aluminum interconnect, **5**-21–22
  cleaning issues, **5**-23
Subtractive integration, damascene copper integration, comparison of, **2**-7–16
Sulfonyloxy imides, **19**-26–27
Sulfuric peroxide (piranha), **5**-20
Sulfuric/oxidizer chemistry, **5**-18–19
Supercritical carbon dioxide cleaning processes, **6**-6–18
  drying, **6**-17–18
  etching, **6**-15–17
  photoresist striping and removal, **6**-6–10
  processing equipment, **6**-18–20
  trace metals and particulates, **6**-11–14
Supercritical carbon dioxide
  semiconductor cleaning and, **6**-1–21
  supercritical fluids, **6**-2–6
Supercritical fluids, **6**-2–6
  characteristic properties, **6**-2–4
  solvation capabilities, **6**-5–6
Surface chemistry
  chemical etching, **21**-8–9
  plasma, **21**-8–10
  polymer deposition, **21**-9
  reactive ion etching, **21**-9
  sputtering, **21**-8
  substrate temperature, **21**-10
Surface cleanings, salicide development and, **10**-14–15
Surface evaporation, **3**-13–15
Surface mobility based deposition, **15**-10–11
Surface preparation, **5**-1–32
  aqueous solutions, **5**-4–7
  bulk organic removal, **5**-18
  cleaning, **5**-21
  complexing, **5**-7–8
  control and monitoring of, **5**-30–32
  copper/low-*k* dual damascene interconnect, **5**-23
  defects and drying, **5**-17
  DI/ozone process, **5**-20–21
  electrochemistry, **5**-2–4
  hydrofluoric acid, **5**-16–17
  oxide etch, **5**-15
  oxide/nitride etch, **5**-18
  photolithography, **5**-18
  photoresist strip, **5**-18
  polysilicon etch, **5**-17
  random defects and, **27**-4
  RCA cleaning process, **5**-2
  reaction mechanism, **5**-19
  SC-1, **5**-7
  selective nitride etch, **5**-17–18
  subtractive aluminum interconnect, **5**-21–22
  sulfuric peroxide, **5**-20
  sulfuric/oxidizer chemistry, **5**-18–19

typical defects, **5**-28–30
wet chemical etching, **5**-15
Surface roughness, **11**-10, **16**-28–29
SOI wafers and, **4**-25
thermal radiative properties and, **11**-26–27
System level redundancy techniques, **31**-19–21
Systematic yield loss, **23**-2–4

T

Ta precursors, high κ dielectric and, **13**-42–43
Ta_2O_5 annealing, **13**-44–45
Ta_2O_5 DRAM applications, **13**-42
Ta_2O_5 integration, **13**-44
Table motion, **17**-7–9
Tandem acceleration, **7**-51–52
Target tracking. *See* compensation process.
Technological advances, semiconductor manufacturing and, **35**-14–18
Temperature dependence, ALD and, **14**-12–13
Temperature differences, process uniformity control and, **11**-65
Temperature measurement
control methods, summary of, **11**-60, 61–62
diffuse reflectance spectroscopy and, **25**-8–9
problems with, **11**-38
RTP systems and, **11**-38–60
combined radiation shield, **11**-57–60
direct couple thermocouple, **11**-57–60
direct thermocouple control, **11**-56–57
hot plates, **11**-51–53
contact sensors, **11**-45, 46
open loop lamp intensity control, **11**-45, 47–51
pyrometry, **11**-38–45
radiation shields, **11**-53–56
thermocouples, **11**-45, 47, 48
Temperature sensor, acoustic wafer, **25**-9–10
Temperature uniformity control, dynamic, **11**-70–71
Temperature, ALD process and, **13**-10–13
Temperature, process state sensors and, **25**-4–10
Terminal effect, **16**-30–31
Terrestrial radiation effects
memory SER sensitivity, **31**-13–15
mitigation techniques, **31**-16–21
semiconductor devices, **31**-10–13
sequential/combinational logic SER sensitivity, **31**-15–16
Terrestrial radiation environment, **31**-2–10
high energy cosmic ray neutrons, **31**-4–6
impurity radioactive decay, **31**-6–8
ions in matter, **31**-2–3
low energy cosmic ray neutrons, **31**-9–10
Terrestrial radiation integrated circuits and, **31**-1–21
Terrestrial radiation mitigation techniques, **31**-16–21
design types, **31**-19
process technology, **31**-17–18
source types, **31**-16
system level redundancy techniques, **31**-19–21
Test yield limits, **26**-16–18
Texture development, **10**-33–34
Thermal Atomic layer deposition reactions, **14**-8–10
catalytic, **14**-11–12
digital control, **14**-16–17
halide-hydride chemistry, **14**-8–9
initiation processes, **14**-17–18
linearity with cycling, **14**-16–17
metal organic chemistry, **14**-9–10
nanolaminates, **14**-19
plasma assisted, **14**-10–11
rapid, **14**-19–20
saturation characteristics, **14**-13–15
throughput calculator, **14**-15–16
temperature dependence, **14**-12–13
Thermal budget issues, **1**-42–44
Thermal characteristics, dislocation free growth, **3**-5–7
Thermal control, **18**-43
Thermal convection-related temperature fluctuations, **3**-12
Thermal decomposition, **13**-75–76
Thermal deformation, **32**-18–22
Thermal equilibrium, defect concentrations, **8**-4–5
Thermal expansion, nickel silicides and, **10**-31–32
Thermal generation based global techniques, **29**-9–10
Thermal management, integrated circuit packaging and, heat sink attachments, **32**-8–9
Thermal processing wafer distortion, **18**-13
Thermal radiation physics, **11**-5–7
basic laws, **11**-5–7
radiative properties, **11**-6–7
Thermal radiative properties, **11**-6–7
calculations of, **11**-7–10
patterns, **11**-10
spectral emissivity, **11**-8–10
surface roughness, **11**-10
integrated optical properties, **11**-7
Kirchhoff's Law, **11**-6
semiconductor wafers, **11**-15–27
absorptivities, **11**-15–27
integrated emissivities, **11**-18–20
silicon spectral emissivity, **11**-15–27
silicon coatings, **11**-20–26
spectral emissivity, **11**-6
surface roughness, **11**-26–27
Thermal response, wafers and, **11**-27–38
heat transfer models, **11**-34–38
heating considerations, **11**-28–32
wafers and, RTP system dynamics, **11**-33–34
Thermally activated reaction, **13**-5–7
Thermocouple control, direct, **11**-56–57
Thermocouples, **11**-45, 47, 48
direct couple, **11**-57–60
Thermodynamics of defects, **8**-5–6
enthalpy of formation, **8**-4
equilibrium of different charge states, **8**-5–6
thermal equilibrium, defect concentrations, **8**-4–5
Thickness distribution, **16**-29
Thin films, **10**-31–32
CTE anisotropy, **10**-32
CVD types, **13**-22–82
Thin gate films, **24**-24–25
optical models, **24**-24–25

Thin silicon oxidation models, **9**-8–11
Threshold voltage, **4**-29–30, **28**-12–14
Throughput calculator, **14**-15–16
Throughput production lot, **18**-19
Throughput rate, **31**-1–2
Throughput
  exposure field, **18**-18
  exposure tool system and, **18**-17–19
  full wafer, **18**-18
Ti silicide process, **10**-7–12
$TiCl_4/NH_3$, **13**-74
Time dependent dielectric breakdown, **30**-23–30
  channel hot carrier injection, **30**-28–30
  complementary models, **30**-27–28
  E-model, **30**-24–25
  I/E model, **30**-25
  mobile-ions/surface inversion, **30**-27–28
Time to failure modeling, **30**-3–6
  USLI failure mechanisms, **30**-11–15
Time to failure statistics, **30**-6–10
  acceleration factor, **30**-10–11
  failure statistics, **30**-10
  lognormal distribution, **30**-7
  Weibull distribution, **30**-8–10
TiN films, **13**-74–75
Titanium, **13**-71–73
Titanium nitride, **11**-96, **13**-73–76
  metal-organic TiN films, **13**-74–75
  plasma treatment cycle, **13**-75–76
  thermal decomposition, **13**-75–76
  $TiCl_4/NH_3$, **13**-74
Titanium silicide, **11**-92–93
Total reflection X-ray fluorescence, **28**-49–51
Trace element radiochemical methods, **28**-45–46
Trace metals, cleaning of, **6**-11–14
Tracking systems, wafer fab logistics and, **33**-3–10
Transconductance, **4**-30–31
Transient effects, **4**-28–29
Transient non uniformity, **11**-65
Transistor structures, **21**-50–51
Transmission electron microscopy, **28**-32–35, **29**-17–18
Transmission measuring, **20**-64–65
Transparent polymer systems, ArF, **19**-41–45
Transport enhanced MOS devices, **1**-50–51
Trench features, **11**-82–83
Triple gate SOI MOSFETS, **4**-41–43
Tungsten CMP, **17**-45
  chemical effects and, **17**-37–39
Tungsten CVD film, **13**-59–67
  barrier layers, **13**-62
  plug formation, **13**-62
  W nucleation process, **13**-62–63
  WCVD nucleation layer, **13**-67
  W-CVD process, **13**-59–61
Tungsten silicides, **13**-68
22 nm nanotechnology, future of
  equipment and processing, **21**-61–62
  etch interactions, **21**-59
  imaging limitation, **21**-59
  silicon, **21**-62–63

U

Ultra large scale integration. *See* USLI.
Ultra low-κ dielectrics, **21**-55–57
Ultra shallow junction formation, **7**-37–40
Ultra thin film measurements, dielectric film metrology and, **9**-22–23
Ultra thin gate dielectrics, **4**-37–39
  gate induced floating body effect, **4**-37
Ultra thin SiO_2, **24**-25–27
Undoped silicon glass, **13**-26–28
Uniformity configurations, **11**-65–69
Uniformity control
  coatings and patterns, **11**-71–74
  dynamic temperature, **11**-70–71
Uniformity gate fabrication integration issues and, **21**-30–32
Uniformity optimization procedures, **11**-69–70
Unintended dopants, **3**-8–9
Unit cell dimensions
  bulk samples, **10**-31
  thin films, **10**-31–32
Unit conversions, **B**-1–2
Univariate statistical process control, **23**-18–19
Unpatterned wafer monitoring, **27**-14–15
USLI (ultra large scale integration) failure mechanisms, **30**-11–15
  contact electromigration, **30**-15–16
  corrosion, **30**-16–18
  cyclic fatigue, **30**-21–23
  electromigration, **30**-11–15
  ICs, Czochralski (CZ) technique, **3**-1
  stress migration, **30**-18–21
  time-dependent dielectric breakdown, **30**-23–30
Utilization metrics, **22**-9

V

Vacancy point defects, **8**-3
Variations requiring compensation, process control and, **23**-10–11
Variations time scale, **23**-11–13
Vector scan system, **20**-22
Very large scale integration (VLSI) integrated circuits, **1**-1
Volume inversion, **4**-32
Volumetric change, nickel silicide and, **10**-19

W

W nucleation process, **13**-62–63
Wafer
  alignment keys, displacement of, **18**-13–14
  backside inspection, **27**-15
  bonding, **4**-7–8
  charging control, **7**-58
  hucking errors, **18**-14
  contamination, **7**-77–79
  cooling, **7**-68–71
  edge

engineering, random defects and, **27**-7
inspection, **27**-24–25
electrical characterization, Ψ-MOSFET, **4**-26–27
emissivity effects, **11**-41–45
fab logistics, **33**-1–64
carriers, **33**-3–10
interbay transport and storage, **33**-10–40
reticle handling metrics, **33**-3
tracking systems, **33**-3–10
wafer metrics, **33**-3
generic mask on, **20**-2–5
metrics, **33**-3
monitoring, unpatterned, **27**-14–15
non-patterned, **17**-25–29
optical properties of, **11**-10–15
polishing patterned, **17**-29–36
preparation, **3**-45–49
chemical etching, **3**-46–47
cleaning, **3**-48–49
edge rounding, **3**-47
lapping/grinding, **3**-47
polishing, **3**-48
slicing, **3**-45–46
semiconductor, **11**-15–27
silicon epitaxial, **3**-49–61
state sensors, **25**-31–47
film thickness and uniformity, **25**-32–44
CMP film removal, **25**-40–41
optical sensors, **25**-32–40
scatterometer, **25**-44–47
thermal response of, **11**-27–38
heat transfer models, **11**-34–38
heating considerations, **11**-28–32
RTP system dynamics, **11**-33–34
Wall deposition sensors, **25**-30–31
Water resist surface effects, **18**-42–43
Wavelength choice, pyrometry and, **11**-40–41
WCVD nucleation layer, **13**-67
Weibull distribution, **30**-8–10
Wet chemical etching, **5**-15
Wet etching of chrome, **20**-43
Writing patterns, **20**-19–39
mask writers, **20**-19–39

X

X-ray analysis, **29**-19
X-ray diffraction, **28**-54–57
X-ray fluorescence, **28**-49–51
total reflection, **28**-49–51
X-ray method, interconnect film thickness and, **24**-37–39
X-ray photoelectron spectroscopy, **28**-51–53
X-ray reflectance, **28**-57–59
X-ray rocking curves, **28**-60
X-ray topography, **28**-57–59

Y

Yield impact prediction/verification, **27**-25–27
Yield management, **27**-1–28
methodology, **27**-8–28
management priority and motivation, **27**-11
process and equipment control, **27**-11–16
in situ contamination monitoring, **27**-12–13
unpatterned wafer monitoring, **27**-14–15
wafer backside inspection, **27**-15
product based defect detection and analysis, **27**-16–25
brightfield inspection, **27**-21–22
darkfield inspection, **27**-22
electron beam inspection, **27**-22–24
macro inspection, **27**-24
wafer edge inspection, **27**-24–25
random defects, **27**-2–7
root cause isolation, **27**-27–28
yield impact prediction/verification, **27**-25–27
Yield modeling, **26**-1–19
cluster analysis, **26**-2–4
general yield model, **26**-7
limits, **26**-7–18
limited yield analysis, **26**-9–11
product sensitivity analysis, **26**-11–16
random defect types, **26**-4–7
test limits, **26**-16–18